# Etude des interactions protéine-protéine et protéine-ligand par bio- et chimie-informatique structurale : Identification de petites molécules bio-actives

Dominique Douguet

# Habilitation à diriger des recherches

Faculté des Sciences – Université de Nice-Sophia Antipolis

**Dominique DOUGUET**

## Etude des interactions protéine-protéine et protéine-ligand par bio- et chimie-informatique structurale : Identification de petites molécules bio-actives

Soutenue le 19 Novembre 2007 devant le jury composé de :

| | |
|---|---|
| Pr Daniel CABROL | Rapporteur |
| Dr Pierre CHARDIN | Président |
| Dr Gilles LABESSE | Examinateur |
| Dr Anne POUPON | Rapporteur |
| Dr Bruno VILLOUTREIX | Rapporteur |

**Institut de Pharmacologie Moléculaire et Cellulaire**
**660, route des Lucioles, 06650 Valbonne**

**DOUGUET Dominique**

Adresse professionnelle:
Centre de Biochimie Structurale
29, rue de Navacelles
F-34090 Montpellier, France
+33 (0)4 67 41 77 01
Email : douguet@cbs.cnrs.fr
Web : www.cbs.cnrs.fr

# Bioinformatique Structurale

***Identification de nouvelles molécules d'intérêt thérapeutique***

## Titres Universitaires

**1998** **Doctorat** en sciences chimiques et biologiques pour la santé de l'Université de Montpellier I (direction par le Pr. G. Grassy du Centre de Biochimie Structurale)
*Etudes de Relations Structure-Activité (QSAR) et développement d'un programme de génération automatique de molécules par algorithme génétique*

**1995** **DEA** *double compétence* Chimie Informatique et Théorique de l'Université de Paris XI

**1992** **Maîtrise de Chimie** de l'Université de Brest

## Expérience Professionnelle

**2007-** Chargée de recherche **CR1 INSERM** affectée au Centre de Biochimie Structurale de Montpellier (U554).

- Identification de molécules bio-actives basée sur le criblage de fragments : intégration des activités de *de novo* 'drug design' dans les études structurales expérimentales par RMN et par diffraction des rayons X.
- Transfert de technologie dans le domaine du *de novo* 'drug design' vers la start-up NovaDecision (programme LEA3D).

**2003-4 Mise à disposition** de 18 mois au Centre de Bioinformatique dirigé par le Pr. Ilya Vakser à l'Université SUNY Stony Brook, NY, USA.

- Analyse et traitement des ressources sur les interactions protéine-protéine.
- Développement d'une base de données annotée de structures de complexes protéine-protéine co-cristallisées afin d'améliorer leur prédiction (http://dockground.bioinformatics.ku.edu).

**2002-** Recrutement en tant que **CR2 INSERM** affectée au Centre de Biochimie Structurale, U554, Montpellier.

- Identification et optimisation de molécules bio-actives par criblage virtuel de chimiothèques et par *de novo* 'drug design' (développement du programme LEA3D).
- Application à des protéines de *Mycobacterium Tuberculosis* (TMPK et MabA) ainsi qu'à certains récepteurs nucléaires (LRH-1, TR4).

**2000-2002** Postdoctorat en **bioinformatique** au Centre de Biochimie Structurale (Montpellier) dans le cadre du programme GENOPOLE Languedoc-Roussillon.

- Développement du serveur web @TOME (@utomatic Threading Optimisation Modeling & Evaluation) permettant l'analyse structurale des séquences, la détection des homologies et la modélisation comparative de la structure 3D de protéines (http://bioserver.cbs.cnrs.fr).

**1999-2000** Postdoctorat en **modélisation moléculaire** dans les laboratoires AVENTIS (anciennement Hoechst Marion Roussel à Romainville).

- Modélisation de protéines impliquées dans les maladies de l'os et en anti-infectieux.
- Création et optimisation de chimiothèques focalisées sur des critères de diversité moléculaire et de 'docking' dans le cadre de projets de pharmacochimie par chimie parallèle et en collaboration avec les biochimistes du 'High Throughput Screening' *in vitro*.

**1996-1999 Doctorat sous contrat CIFRE** avec les laboratoires GALDERMA R&D (filiale pharmaceutique des groupes L'Oréal et Nestlé, Sophia-Antipolis) et le Centre de Biochimie Structurale (Montpellier).

- Développement d'un programme de conception et d'optimisation de structures de molécules assisté par ordinateur et utilisant les concepts d'évolution Darwinienne (programme LEA).
- Etudes des relations structure-activité (QSAR) de rétinoïdes, agonistes des récepteurs nucléaires RARs.

**1994-1995** Stage de DEA au laboratoire de Chimie Physique des Matériaux Amorphes sous la direction du Pr. A. Fuchs (Université Paris XI, Orsay).

- Etude de l'adsorption d'argon et d'azote dans la zéolite silicalite-1 par simulation numérique grand canonique de Monte-Carlo.
- Intégration des algorithmes 'Reaction Field' et des sommes d'Ewald dans le calcul électrostatique des interactions afin d'améliorer la corrélation avec les résultats expérimentaux.

**Publications**

1 Gao Y., **Douguet D**., Tovchigrechko A. and Vakser I.A., DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking, *Proteins*, **2007**, Special Issue on the Critical Assessment of PRedicted Interactions 2006 (in press).

2 **Douguet D**., Chen H.C., Tovchigrechko A. and Vakser I.A., DOCKGROUND resource for studying protein-protein interfaces, *Bioinformatics*, **2006**, *22*(21), 2612-2618.

3 Drin G., **Douguet D**. and Scarlata S., The Pleckstrin Homology Domain of Phospholipase Cβ Transmits Enzymatic Activation through Modulation of the Membrane-Domain Orientation, *Biochemistry*, **2006**, *45*(18), 5712-24.

4 **Douguet D**., Munier-Lehmann H., Labesse G. and Pochet S., LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design, *J. Med. Chem*., **2005**, *48*, 2457-2468.

5 Munier-Lehmann H., Pochet S., Dugue L., Dutruel O., Labesse G., **Douguet D**., Design of Mycobacterium tuberculosis thymidine monophosphate kinase inhibitors. *Nucleosides, Nucleotides & Nucleic Acids*, **2003**, *22* (5-8), 801-804.

6 Labesse G., Bucurenci N., **Douguet D**., Sakamoto H., Landais S., Gagyi C., Gilles A.M. and Bârzu O., Comparative modelling and immunochemical reactivity of *Escherichia coli* UMP kinase, *Biochem. Biophys. Res. Commun.* , **2002**, *294* (1), 173-179.

7 Labesse G., **Douguet D**., Assairi L. and Gilles A.M., Diacylglyceride kinases, sphingosine kinases and NAD kinases are distantly related to 6-phosphofructokinases, *TIBS*, **2002**, *27* (6), 273-275.

8 Pochet S., Dugue L., **Douguet D**., Labesse G. and Munier-Lehmann H., Nucleoside Analogs as Inhibitors of Thymidylate Kinases: Possible Therapeutical Applications, *Chem. Bio. Chem*., **2001**, *1*, 108-110.

9 **Douguet D**., Bolla J.-M., Munier-Lehmann H. and Labesse G., From sequence to structure to function: a case study, *Enzyme and Microbial Technology*, **2001**, *30* (3), 289-294.

10 **Douguet D**. and Labesse G., Easier threading through web-based comparisons and cross-validations, *Bioinformatics*, **2001**, *17*, 752-753.

11 **Douguet D**., Thoreau E. and Grassy G., LEA (Ligand by Evolutionary Algorithm): A Genetic Algorithm for the Automated Generation of Small Organic Molecules, *J. Comput.-Aided Mol. Design*, **2000**, *14*, 449-466.

12 **Douguet D**., Thoreau E. and Grassy G., A Quantitative Structure-Activity Relationships Studies of RAR α, β, γ Retinoid Agonists, *Quant. Struct.-Act. Relat*., **1999**, *18*, 107-123.

13 **Douguet D**., Pellenq R. and Fuchs A., The Adsorption of Argon and Nitrogen in Silicalite-1 Zeolite : A Grand Canonical Monte-Carlo Study, *Molecular Simulation*, **1996**, *17*, 255-288.

## Chapitre d'ouvrage

Cohen-Gonsaud M., Catherinot V., Labesse G., **Douguet D**., From molecular modeling to drug design, *Practical Bioinformatics*, Bujnicki, Janusz M (ed), Springer-Verlag, New York, **2004**; pp 35-72.

## Brevets

Quémard A., Labesse G., Daffé M., Ducasse S., Cohen-Gonsaud M., **Douguet D**., Marrakchi H., Use of the Protein MABA (FABG1) of *Mycobacterium tuberculosis* for designing and screening antibiotics, **Patent EP1490491** (filed: March 28th 2003).
**Ce brevet est en cours de valorisation.**

Munier-Lehmann H., Douguet D., Labesse G., Pochet S., New aryl pyrimidyl compounds, pharmaceutical compositions comprising them, their use as antimicrobial agents, **Demande PCT/ EP2005/012346** (filed: November 4th 2005).

## Proceeding

Labesse G., **Douguet D**., Gracy J., Pons J.-L. and Chiche L., Stepwise modelling optimization from sequence to quaternary structure, *Protein Science*: Proceedings of 4th European Symposium of The Protein Society, **2001**, *10*, 140.

**Encadrement**

| | |
|---|---|
| Nov. 2005-Jui. 2006 | Gaëlle Guinefort, étudiante en Master Recherche 'Biologie-Santé' 2ème année de l'Université Montpellier I et II |
| Janvier-Avril 2005 | Gaël Lagrange, ingénieur en bioinformatique en CDD au Centre de Biochimie Structurale |
| Janvier-Avril 2005 | Aurélien De Brix, étudiant en Master 'Biologie-Santé' 1ère année de l'Université Montpellier I et II |
| Nov. 2003-Fév. 2004 | Huei-Chi Chen, étudiante à l'Université de Stony Brook, NY, US (niveau équivalent au Master 1ère année) |
| Janvier-Juin 2003 | Julien Viaud, étudiant au DEA 'Interface Chimie Biologie' de Montpellier I |
| Janvier-Juin 2003 | Guillaume Poncet, étudiant au DEA 'Interface Chimie Biologie' de Montpellier I |
| Janvier-Juin 2002 | Grégori Gerebtzoff, étudiant au DEA 'Interface Chimie Biologie' de Montpellier I |
| Mars-Juin 2001 | Emma Ribes, étudiante en Maîtrise de Biochimie de Montpellier II |

**Enseignement**

| | |
|---|---|
| Novembre 2005<br>Décembre 2004 | Cours sur les interactions protéine-ligand dans le cadre de l'option Biologie Structurale du Master 'Biologie-Santé' (1ère année) de l'Université Montpellier I et II |
| Avril 2003 | Travaux pratiques : "Analyse de séquences et construction de modèles tridimensionnels de protéines" au DEA 'Biologie-Santé' de l'Université Montpellier II |
| Avril 2003<br>Mars 2002 | Cours sur le 'drug design' (docking et conception de ligands) au DESS de Bioinformatique de l'Université Montpellier II |
| Décembre 2002 | Cours sur le 'drug design' (docking et conception de ligands) au DEA 'Interface Chimie Biologie' de l'Université Montpellier I |
| Mars 2002 | Travaux pratiques : "Modélisation moléculaire: exemples d'assemblage ligand-protéine" au DEA 'Biologie-Santé' de l'Université Montpellier II |
| 14-18 Janvier 2002 | Animatrice dans le cadre de la formation CNRS "Exploitation des séquences & Modélisation moléculaire" (Paris) |
| Novembre 2001 | Cours "Eléments de bioinformatique" au DEA 'Interface Chimie Biologie' de l'Université Montpellier I |
| Novembre 2001 | Travaux pratiques : "Modélisation moléculaire: exemples d'assemblage ligand-protéine" au DEA 'Interface Chimie Biologie' de l'Université Montpellier I |
| 5-6 octobre 2001 | Travaux pratiques sur l'analyse des séquences et la construction de modèles tridimensionnels de protéines dans le cadre des ateliers "Bio-Informatique Structurale" à Montpellier proposés par la GENOPOLE Languedoc-Roussillon |
| Octobre 2001 | Cours "Eléments de bioinformatique" à l'IUP 'Ingénierie de la Santé' de l'Université Montpellier I |

Dès le lycée, j'ai eu la chance de suivre un enseignement en informatique. A l'époque (1987-1990), l'accès aux ordinateurs et leur utilisation comme outil de calcul via la programmation était l'apanage de quelque rares établissements de l'enseignement secondaire. Les ordinateurs 'Apricot' (microprocesseur à 5 Mhz et une RAM de 512 Ko !) font maintenant partis de l'histoire… Mon goût pour la chimie était cependant plus prononcé et, très tôt, j'ai décidé d'étudier cette discipline via un parcours universitaire à l'Université de Bretagne Occidentale à Brest. J'y ai obtenu un DEUG A puis une licence et une maîtrise de chimie organique. Mon but était d'obtenir un doctorat. J'ai atteint cet objectif grâce à mon intérêt pour l'informatique et en m'orientant vers un DEA double compétence chimie et informatique que proposait l'Université Paris XI à Orsay. Quelques mois plus tard, je fus recrutée par le CIRD GALDERMA (filiale pharmaceutique des groupes l'Oréal et Nestlé implantée à Sophia-Antipolis) pour réaliser une thèse dans le cadre d'un contrat CIFRE. Mes connaissances dans le domaine de la chimie et de l'informatique furent dès lors dédiées à des applications dans le domaine de la biologie.

Ma thèse de doctorat fut donc réalisée sous la cotutelle de la société Galderma et du Centre de Biochimie Structurale (Montpellier). Ces travaux concernaient l'étude de petites molécules modulatrices de l'activité des récepteurs nucléaires à l'acide rétinoïque (RARs).

L'analyse des relations structure-affinité (SAR) d'agonistes des RARs nous a conduit à l'élaboration de trois modèles statistiques par analyse linéaire et non linéaire d'une base de 140 rétinoïdes d'affinité connue sur les récepteurs RAR $\alpha$, $\beta$ et $\gamma$. Ces modèles appartiennent à la même famille mais ils ont mis en évidence des caractéristiques physico-chimiques propres à chaque récepteur. Les modèles de type non linéaire obtenus par la méthode du 'variable mapping' se sont révélés plus particulièrement intéressants car ils constituaient une analyse décisionnelle adaptée aux spécifications d'un algorithme génétique, sujet de la seconde partie de ma thèse. En effet, l'intérêt des études SAR, en dehors de leur caractère explicatif, tient surtout de leur pouvoir prédictif. La difficulté est alors d'effectuer une démarche inverse à celle de l'analyse SAR afin de proposer des nouvelles molécules en adéquation avec les propriétés physico-chimiques mises en avant par les modèles.

Ce travail m'a donc permis d'acquérir une expérience dans le domaine de la **chémoinformatique** par des études de type **relations structure-activité** (Douguet, et al., 1999) ainsi que par la création d'un programme de ***de novo* 'drug design'** (Douguet, et al., 2000)**.** LEA (Ligand by Evolutionary Algorithm) est un programme de conception et d'optimisation de structures de molécules assisté par ordinateur. L'algorithme génétique, d'inspiration Darwinienne, applique les principes de la sélection naturelle afin d'introduire des concepts d'évolution structurale sous contraintes. Les structures moléculaires évoluent au cours d'un certain nombre de générations jusqu'à l'apparition de molécules adaptées aux contraintes spécifiées. Ces dernières peuvent être issues d'une analyse des relations structure-activité (SAR) telles que celles décrites précédemment.

Par la suite, au sein de la société AVENTIS (anciennement Hoechst Marion Roussel à Romainville (93)), j'ai mis en œuvre des méthodologies de **criblages virtuels de chimiothèques** à haut débit dans le cadre de recherche de ligands pour des cibles thérapeutiques variées aussi bien en pathologie osseuse qu'en thérapeutique anti-infectieuse. Mon travail de modélisation impliquait la modélisation des protéines étudiées et la préparation des banques de petites molécules que l'on souhaitait cribler. Divers critères de sélection (diversité moléculaire, affinité prédite, recherche par pharmacophore,…) permettaient de sélectionner des ligands potentiels. Ainsi, en collaboration avec les chimistes

et les biologistes du service 'HTS' *in vitro*, nous avons permis la synthèse de nouveaux inhibiteurs, l'optimisation de molécules synthétisées par chimie parallèle et la sélection de sous-ensembles de molécules à tester *in vitro*. J'ai également participé à l'élaboration d'expériences de mutagenèse dirigée pour comprendre le mécanisme enzymatique d'une cible protéique. J'ai étudié plus particulièrement 5 protéines : une impliquée dans les maladies de l'os, une en oncologie et trois autres en anti-infectieux.

Suite à mon passage dans ces départements de R&D, j'ai réintégré le milieu académique où j'ai eu l'opportunité d'étendre mes connaissances de la modélisation par homologie de la structure des protéines et de leurs interactions avec des ligands ou avec d'autres protéines. Ces travaux vont être décrits dans le document qui va suivre.

# Habilitation à diriger des recherches

Faculté des Sciences – Université de Nice-Sophia Antipolis

**Dominique DOUGUET**

## Etude des interactions protéine-protéine et protéine-ligand par bio- et chimie-informatique structurale : Identification de petites molécules bio-actives

Soutenue le 19 Novembre 2007 devant le jury composé de :

| | |
|---|---|
| Pr Daniel CABROL | Rapporteur |
| Dr Pierre CHARDIN | Président |
| Dr Gilles LABESSE | Examinateur |
| Dr Anne POUPON | Rapporteur |
| Dr Bruno VILLOUTREIX | Rapporteur |

**Institut de Pharmacologie Moléculaire et Cellulaire**
**660, route des Lucioles, 06650 Valbonne**

**Sommaire**

# I - Introduction

Les projets de séquençage de génomes entiers délivrent un grand flot d'informations qui doivent être analysées afin de relier génotype et phénotype à différents degrés de précision. L'annotation génomique est l'une des premières tâches mais la détermination précise des fonctions voire des structures tridimensionnelles des protéines apparaît maintenant essentielle. Il est également probable qu'un nombre de plus en plus important de protéines, de structures 3D inconnues, vont être identifiées comme étant de nouvelles cibles thérapeutiques (figure 0 ; (Russ and Lampel, 2005)). Connaître leur structure tridimensionnelle est un élément déterminant pour la compréhension fine de leur mécanisme d'action et indispensable pour le développement d'approches thérapeutiques rationnelles. Ainsi, l'identification et l'analyse structurale des sites de fixation de leurs ligands (protéine ou petite molécule) permettront d'envisager la modulation de leur fonction biologique. Ces études principalement expérimentales peuvent être accélérées par l'utilisation d'outils bioinformatiques.

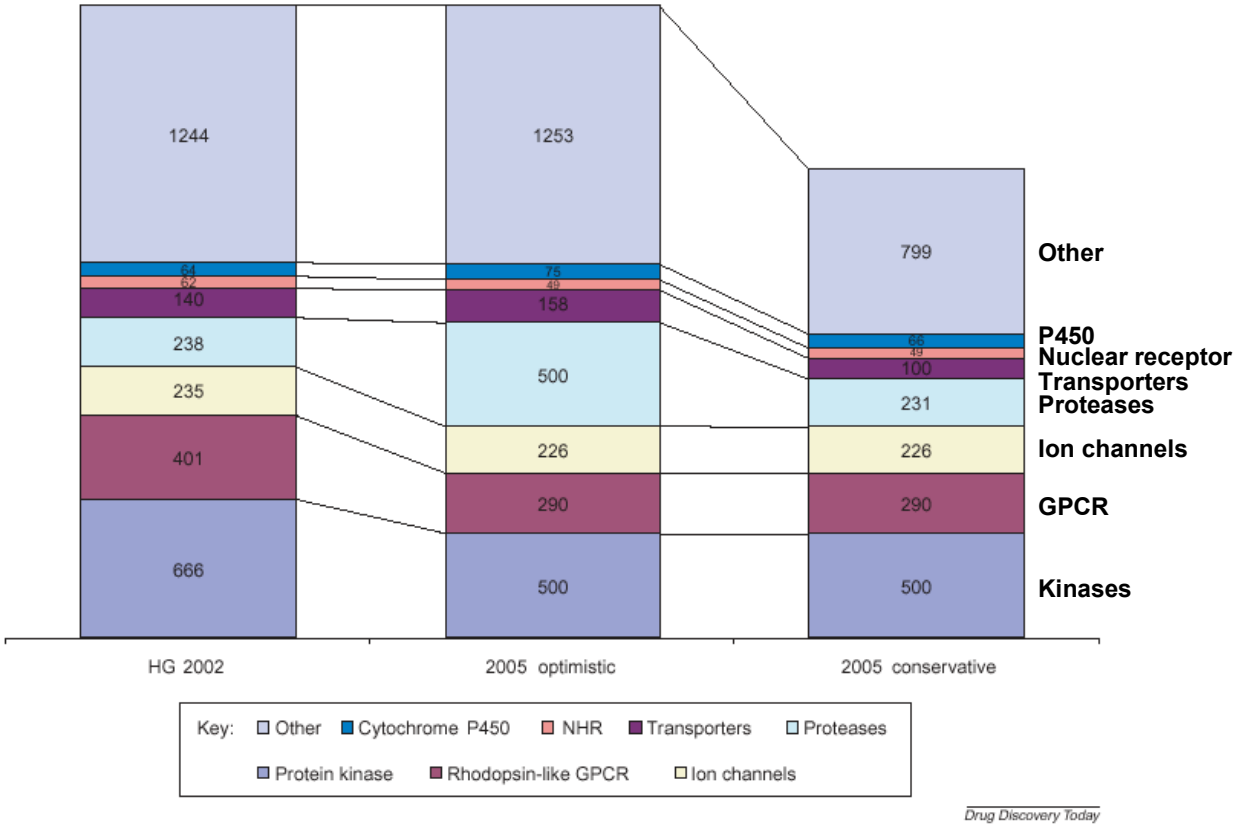



**Figure 0.** Prediction du nombre de cibles 'druggables' d'après Russ and Lampel (Russ and Lampel, 2005).

La bioinformatique est un domaine à forte interdisciplinarité qui utilise des techniques et des concepts de l'informatique, la génétique, la biochimie, la chimie, des mathématiques, des statistiques, etc. … Les recherches dans cette discipline consistent à développer des méthodes pour la collection, le stockage, l'organisation, la visualisation et l'analyse des données biologiques. Il s'en suit le développement de méthodes de prédiction propres à chaque problématique. La bioinformatique structurale est une des composantes de la

bioinformatique qui se réfère principalement à l'étude des macromolécules et plus particulièrement à celle des protéines. Différents aspects de la bioinformatique structurale seront abordés dans ce document : la modélisation par homologie de la structure des protéines, une incursion à l'étude des interactions entre protéines et l'identification/conception de ligands basée sur la connaissance de la structure tridimensionnelle d'une protéine. Cette dernière activité que l'on nomme encore modélisation moléculaire peut être dorénavant considérée comme une extension naturelle du domaine d'étude de la bioinformatique structurale aux interactions entre protéines et petites molécules modulatrices de leur activité.

## II - Modélisation comparative : le serveur @TOME

Entre 2000 et 2002, j'ai travaillé à l'automatisation d'une procédure de modélisation par homologie de la structure des protéines dans le cadre du programme GENOPOLE Languedoc-Roussillon. Cette opportunité m'a été offerte par Gilles Labesse, à l'époque jeune entrant au laboratoire et spécialiste dans le domaine de la modélisation par homologie, notamment à basse identité de séquence.

La modélisation par homologie permet d'obtenir un modèle tridimensionnel d'une protéine lorsque sa structure n'a pas été déterminée expérimentalement (Sali and Blundell, 1993). Elle est basée sur l'observation que les séquences protéiques évoluent plus rapidement que leur structure 3D. Ceci permet d'utiliser comme structure exemple (ou support) une protéine homologue dont la structure à été déterminée expérimentalement (Figure 1 ; (Chothia and Lesk, 1986)).

Si le nombre de structures protéiques déterminées expérimentalement est maintenant conséquent avec plus de 40 000 entrées PDB et environ 6000 nouvelles entrées par an, les méthodes expérimentales de diffraction des rayons X et de résonance magnétique nucléaire (RMN) restent complexes et lentes comparées aux méthodes de séquençage (problèmes d'expression, de solubilité, de purification et de cristallisation des protéines). La détermination de la structure de l'ensemble des protéines exprimées est inenvisageable. Par contre, l'apport de la 'génomique structurale' est très attendue dans la détermination de structures représentatives de l'ensemble des familles protéiques, desquelles la structure des homologues pourra être déduite (Todd, et al., 2005).

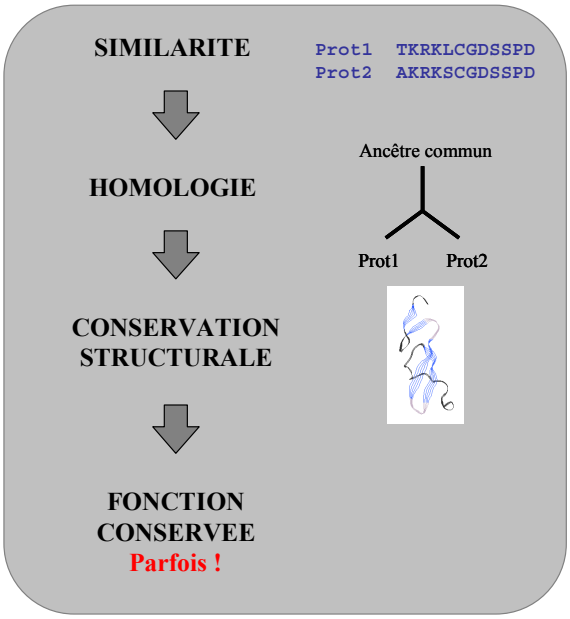



**Figure 1.** Une identité de séquence élevée (>30%) montre une relation structurale claire, et, à une homologie structurale correspond généralement une homologie fonctionnelle, mais les exceptions existent (Devos and Valencia, 2000).

En effet, il existe une certaine redondance parmi les protéines déterminées expérimentalement. Ainsi, sur les 40000 structures expérimentales actuelles, les statistiques de la PDB (juillet 2007) font état de 9000 familles de protéines ne partageant pas plus de 30% d'identité de séquence (6124 selon PISCES (Dunbrack, 1999) et ~17000 à 95% d'identité de séquence selon la PDB). Néanmoins, la modélisation entièrement automatique d'une protéine n'est raisonnable qu'au dessus de 30% d'identité de séquence (Baker and Sali, 2001).

La modélisation par homologie comprend quatre étapes : une étape de reconnaissance du repliement avec la détermination de la structure support expérimentale associée, une étape d'alignement entre séquences, une étape de construction du modèle 3D et, enfin, une étape d'évaluation de la qualité globale du modèle (Figure 2).

L'automatisation de la modélisation par homologie et son accès libre via Internet étaient encore en émergence comme l'indique la liste des serveurs qui ont participé au CASP4 (Critical Assessment of techniques for Protein Structure Prediction) en 2000 (http://predictioncenter.org/casp4). A l'époque, ils étaient surtout spécialisés dans l'une des 4 étapes mentionnées ci-dessus et étaient plus particulièrement dédiés à la prédiction de la structure secondaire et/ou du repliement (exemples de PSI-BLAST (Altschul, et al., 1997) pour l'alignement des séquences, de mGenTHREADER pour la sélection du support (Jones, et al., 1999) …).

L'automatisation complète de la procédure de modélisation par homologie devenait nécessaire au regard de la quantité de données fournie par le séquençage des génomes et la multiplicité des outils à utiliser pour finaliser les modèles protéiques. Ce travail pouvait prendre jusqu'à plusieurs jours pour une seule cible. Au-delà de faciliter le travail du modélisateur, il semblait indispensable de rendre accessible cette procédure aux non experts.

C'est dans la volonté de créer un système intégré capable de réaliser la modélisation complète allant de la recherche de la structure parente la plus homologue à la construction du modèle que l'équipe de bioinformatique du CBS a rejoint le programme GENOPOLE et réalisé le serveur @TOME (@utomatic Threading Optimisation Modeling & Evaluation) accessible à l'adresse http://bioserver.cbs.cnrs.fr ; (Douguet and Labesse, 2001)).

Chacune des quatre étapes citées plus haut peut être réalisée à partir de nombreux outils parfois accessibles via Internet. Un aperçu des outils disponibles en 2007 est donné à la fin de ce chapitre.

## Etape 1 : Reconnaissance du repliement

La ou les structures supports peuvent être identifiées par comparaison des séquences (PSI-BLAST (Altschul, et al., 1997)) ou bien par des méthodes de comparaison séquence/structure de type enfilage ou 'threading' qui peut parfois révéler des relations éloignées entre séquences (Torda, 1997). Le 'threading' consiste à apposer une séquence d'acides aminés sur la structure 3D d'une autre protéine. L'ensemble des repliements connus (structures 3D) peut être criblé de cette manière afin de détecter le plus compatible avec la séquence en acides aminés. Contrairement à l'alignement de séquences classique, l'évaluation, ici, porte sur la compatibilité séquence-structure : des acides aminés éloignés en séquence peuvent se retrouver très proches au niveau 3D. Cette compatibilité entre résidus 'qui se voient' sert d'évaluation.

A ce stade, nous avons développé un méta-serveur qui interroge 6 serveurs externes mis en avant lors du CASP4/CAFASP2 en 2000 : mGenThreader (Jones, et al., 1999), 3D-PSSM (Kelley, et al., 2000), FUGUE (Shi, et al., 2001), SAM-T99 (Karplus, et al., 1998), Jpred2 (Cuff and Barton, 2000) et PDB-BLAST. Un méta-serveur équivalent au nôtre est

apparu au même moment, montrant bien une tendance à l'utilisation de méthodes consensuelles (Bujnicki, et al., 2001). Les résultats du CASP5/CAFASP3 (2003) qui eut lieu quelques années après le début de mon travail soulignèrent très bien la supériorité des méthodes de type méta-serveur sur chacune des méthodes prises indépendamment :

"The performance of the best automated meta-predictors was roughly 30% higher than that of the best independent server. More significantly, the performance of the best automated meta-predictors was comparable with that of the best 5-10 human CASP predictors" (Fischer, et al., 2003).

Ainsi en 2000, notre objectif était de rassembler et de comparer en une seule page web les résultats de différentes méthodes et de privilégier le consensus quant au choix de la structure support. Le méta-serveur que nous avons développé donne la priorité aux supports bien classés dans les serveurs d'origine. Le score 'consensus' est ensuite pondéré par le taux d'identité de séquence, par le pourcentage de recouvrement (ou étendue de l'alignement) et par le score TITO (décrit ci-après à l'étape 2) qui peut privilégier un alignement par rapport à un autre. Au final, un menu déroulant affiche le classement des paires 'support-alignement' par ordre décroissant de priorité.

La figure 3a et 3b présente un exemple récent issu de l'étape 1. L'objectif est de modéliser le domaine catalytique de la phospholipase $A_2$ sécrétée humaine de groupe III dont la structure expérimentale n'a pas encore été déterminée (code uniprot Q9NZ20, segment 151 à 248).

Les résultats montrent sans ambiguïté la compatibilité de cette séquence avec la structure support 1POC avec un taux d'identité de séquence supérieur à 40% (figure 3c et 3d). La protéine 1POC, issue du venin d'abeille (European honeybee (*Apis mellifera*)), appartient également à la superfamille des phospholipases $A_2$.

## Etape 2 : Optimisation de l'alignement des séquences

Le point crucial dans la modélisation par homologie reste l'alignement des séquences. Elle est d'autant plus délicate que l'identité de séquence est faible. Dans la version actuelle du serveur @TOME, l'optimisation, si elle a lieu, est manuelle. Mais, au-delà de 30% d'identité de séquence, les alignements fournis par le méta-serveur peuvent être utilisés dans l'état. Dès l'étape 1, nous proposons une classification des alignements pour chaque structure support basée sur le score TITO. TITO (Tool for Incremental Threading Optimization) est un 'threader' qui calcul un score de compatibilité 1D/3D (Labesse and Mornon, 1998). Ce score permet d'évaluer d'une manière identique l'ensemble des résultats issus du méta-serveur. D'un clic de la souris, on choisit l'alignement optimal qui sera soumis à l'étape 3 (figure 3C et figure 4).

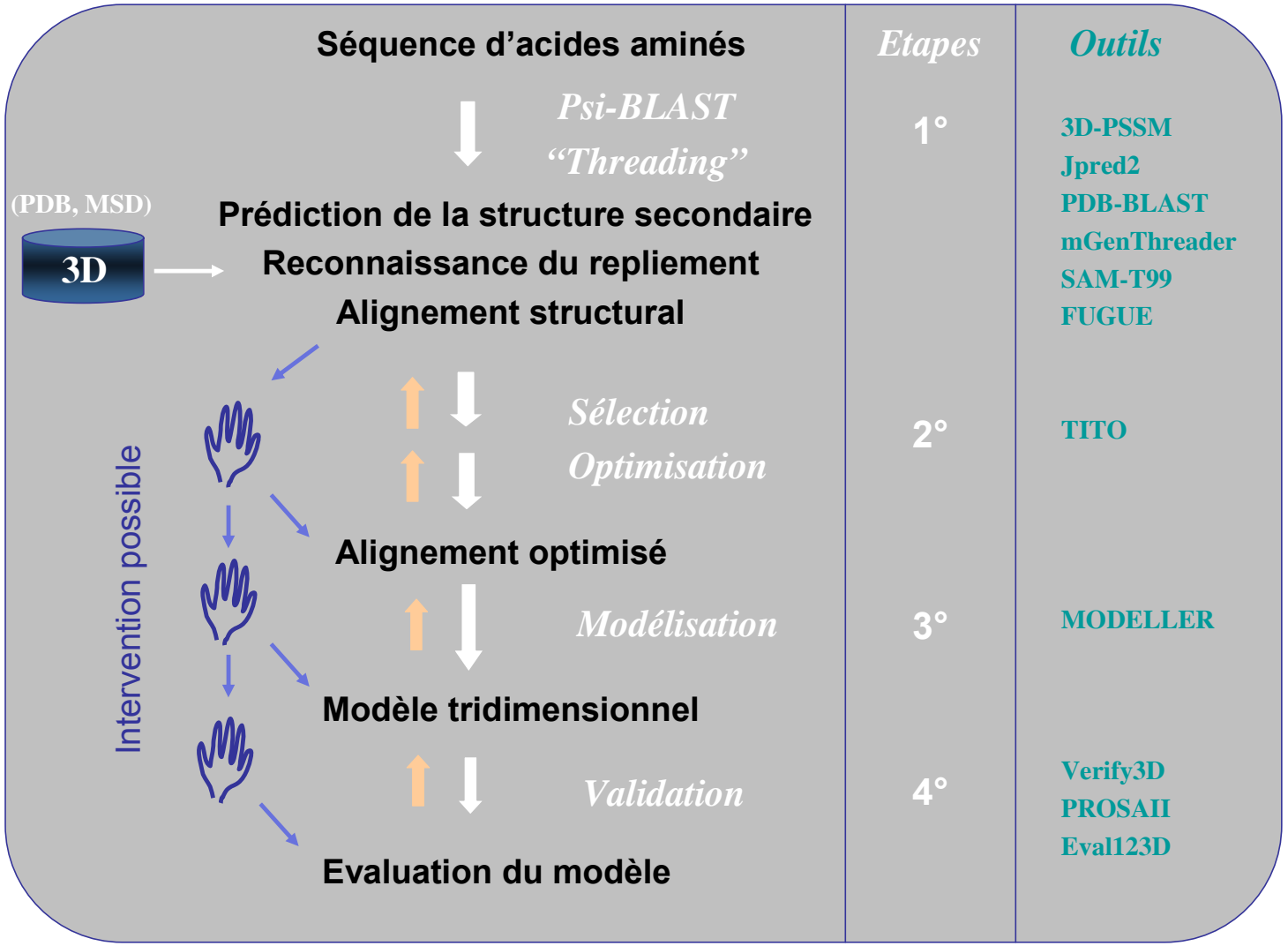
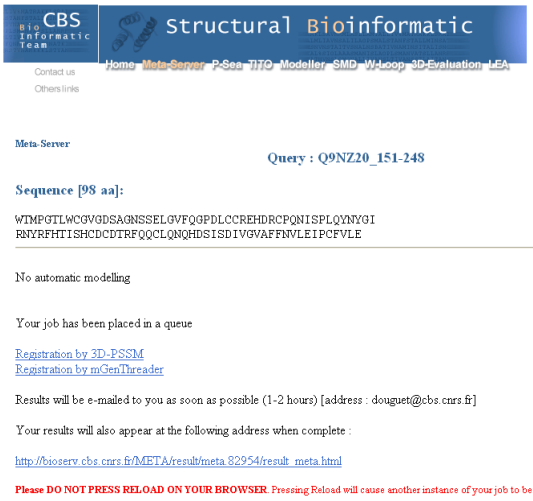



**Figure 2.** Procédure de modélisation par homologie en quatre étapes : 1° Reconnaissance du repliement et choix de la structure support, 2° Alignement des séquences, 3° Construction du modèle et 4° Evaluation globale du modèle.

**Figure 3.** Procédure de modélisation par homologie de la phospholipase $A_2$ sécrétée humaine de groupe III, étape 1° : a) Page d'accueil et formulaire d'entrée du méta-serveur. Le segment 151-248 est extrait de la séquence SWISSPROT Q9NZ20 (509 acides aminés) puis soumis au serveur. Ce segment correspond au site catalytique de la phospholipase. La partie N-ter (1-150) et la partie C-ter (249-509) ne présentent pas d'homologie claire avec une quelconque structure expérimentale. b) Résultat de l'enregistrement. c) Résultats de la reconnaissance du repliement. La structure support 1POC est clairement mise en avant par les serveurs, par la compatibilité 1D/3D par TITO et enfin par le classement par priorité (cercles rouges). d) Données PDB sur 1POC.

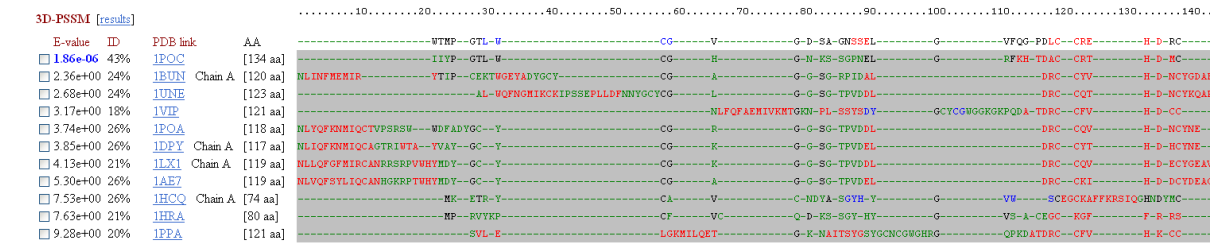
a)

b)

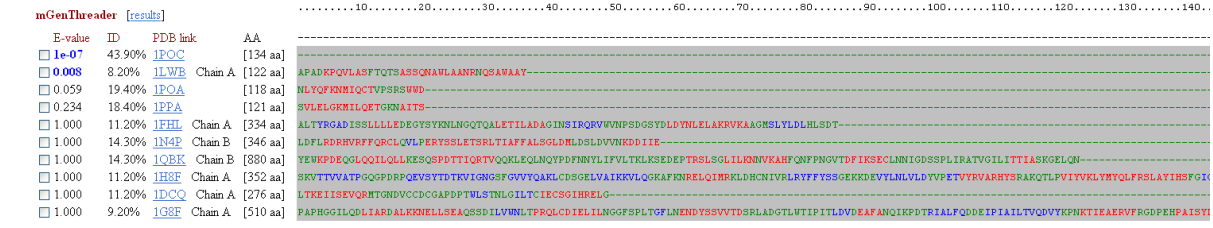



c)

A blue score means a significant Hit




**Citation**: Kelley LA, MacCallum RM & Sternberg MJE (2000). Enhanced Genome Annotation using Structural Profiles in the Program 3D-PSSM. *J. Mol. Biol.* **299**(2), 501-522.




**Citation**: Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J.Mol Biol.* **287**: 797-815.

**Summary:**

pdb 1poc

| Server | Chain | Rank in Server | Server Score | TITO Score |
|---|---|---|---|---|
| 3D-PSSM | - | 1 | 1.86e-06 | -34635 |
| mGenThreader | - | 1 | 1e-07 | -34818 |

Top 10 i.d.

1poc = 43.90% | 1lccq = 26% | 1ae7 = 26% | 1dpy = 26% | 1poa = 26% | 1une = 24% | 1bun = 24% | 1hra = 21% | 1lx1 = 21% | 1ppa = 20%

Best TITO scores of the Top 10 PDB templates

1fhl = -43291 | 1n4p = -41000 | 1poc = -34818 | 1ae7 = -34340 | 1poa = -25611 | 1ppa = -23292 | 1une = -22339 | 1dcq = -22093 | 1dpy = -19724 | 1vip = -18201

*!! Caution must be exercise when a positive TITO score results*

Top 5 3D-PSSM Hits

1poc = 1 | 1bun = 2 | 1une = 3 | 1vip = 4 | 1poa = 5

Top 5 mGenThreader Hits

1poc = 1 | 1lwb = 2 | 1poa = 3 | 1ppa = 4 | 1fhl = 5

Select a PDB structure and click the 'Compute' button to access the **Automatic Modelling of Q9NZ20_151-248 3D Structure** (via a TITO step) : mGenThreader PDB:1poc:- TITO score = -34818

The above first selected PDB (in selection box) seems to be the best template (by consensus evaluation)

Compute

d)

**1poc** DOI 10.2210/pdb1poc/pdb

Red - Derived Information

| | |
|---|---|
| *Title* | CRYSTAL STRUCTURE OF BEE-VENOM PHOSPHOLIPASE A2 IN A COMPLEX WITH A TRANSITION-STATE ANALOGUE |
| *Authors* | Scott, D.L., Otwinowski, Z., Sigler, P.B. |
| *Primary Citation* | **Scott, D.L., Otwinowski, Z., Gelb, M.H., Sigler, P.B.** Crystal structure of bee-venom phospholipase A2 in a complex with a transition-state analogue. *Science v250 pp.1563-1566 , 1990* [Abstract] |
| *History* | Deposition 1992-09-07 Release 1993-10-31 |
| *Experimental Method* | Type X-RAY DIFFRACTION Data N/A |

| *Parameters* | Resolution[Å] | R-Value | R-Free | Space Group |
|---|---|---|---|---|
| | 2.00 | 0.192 (obs.) | n/a | $I\,4_1\,2\,2$ |

| *Unit Cell* | | | | | | |
|---|---|---|---|---|---|---|
| Length [Å] | a | 89.50 | b | 89.50 | c | 132.50 |
| Angles [°] | alpha | 90.00 | beta | 90.00 | gamma | 90.00 |

*Molecular Description Asymmetric Unit*: Polymer: 1 Molecule: PHOSPHOLIPASE A2 Chains: _ EC no.: 3.1.1.4

*Classification*: **Hydrolase**

*Source*: Polymer: 1 Scientific Name: Synthetic construct

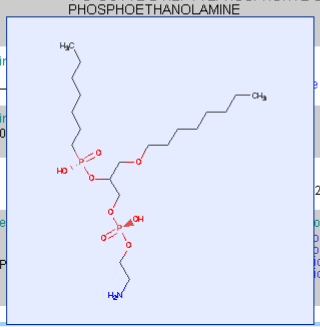
| *Chemical Component* | Identifier | Name | Formula | Drug Similarity | Hapten Similarity | Ligand Structure | Ligand Interaction |
|---|---|---|---|---|---|---|---|
| | GEL | 1-O-OCTYL-2-HEPTYLPHOSPHONYL-SN-GLYCERO-3-PHOSPHOETHANOLAMINE | $C_{20}H_{46}NO_8P_2$ | | | [ View ] | [ View ] |
| | CA | | $Ca^{2+}$ | | | [ View ] | [ View ] |

| *SCOP Classification* (version 1.71) | Domain | Superfamily | Family | Domain | Species |
|---|---|---|---|---|---|
| | d1poc_ | e A2, PLA2 Phospholipase A2, PLA2 | Insect phospholipase A2 | Phospholipase A2 | European honeybee (Apis mellifera) |

| *CATH Classification* (version v3.0.0) | Domain | Architecture | Topology | Homology |
|---|---|---|---|---|
| | 1poc00 | Up-down Bundle | Phospholipase A2 | Phospholipase A2 |

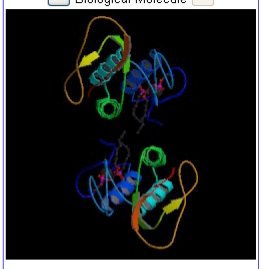
| *PFAM Classification* | Chain | Description | Type | Clan ID |
|---|---|---|---|---|
| | 2 | Phospholipase A2 | Family | n/a |

| *GO Terms* | Polyme | Molecular Function | Biological Process | Cellular Component |
|---|---|---|---|---|
| | PHOSP | olipase A2 activity<br>olipase A2 activity<br>ion binding<br>ion binding | • phospholipid metabolic process<br>• phospholipid metabolic process<br>• lipid catabolic process<br>• lipid catabolic process | • extracellular region<br>• extracellular region |



**Images and Visualization**

<< Biological Molecule >>



**Display Options**

KiNG
Jmol
WebMol
MBT SimpleViewer*
MBT Protein Workshop
QuickPDB
All Images

* Capable of displaying biological molecules.

**Etape 3 : Construction du modèle tridimensionnel**

La construction du modèle par homologie implique la modélisation séquentielle ou simultanée du cœur commun de la protéine, des boucles et des chaînes latérales. Ces trois éléments correspondent, respectivement, aux acides aminés conservés et appareillés dans l'alignement, aux insertions/délétions et aux acides aminés mutés. Plusieurs méthodes existent pour réaliser cette étape : SWISS-MODEL (combinaison de ProModII (Schwede, et al., 2003) et Gromos96), MODELLER (Sali and Blundell, 1993), COMPOSER (Topham, et al., 1990) ou Geno3D (Combet, et al., 2002).

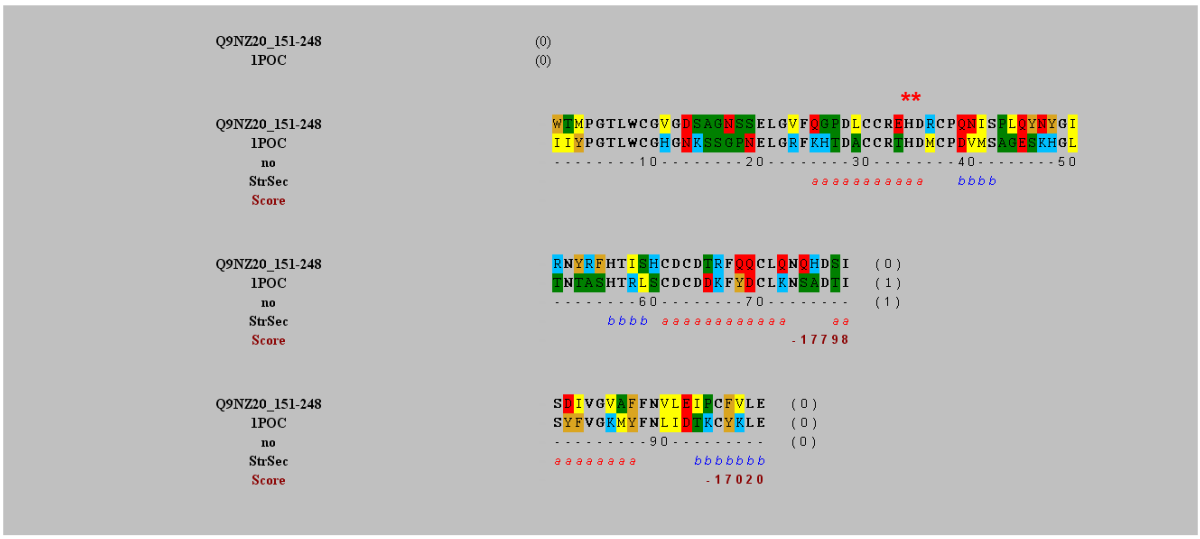
MODELLER est le programme le plus couramment utilisé en modélisation par homologie de par ses performances reconnues. Il construit ses modèles à partir des contraintes spatiales calculées sur la ou les structures supports fournies dans l'alignement (longueur des liaisons, angles de valence et autres préférences statistiques qui constituent la fonction objective à minimiser). Le mode automatique tel qu'il a été définis sur @TOME prépare les fichiers nécessaires à MODELLER pour générer 3 modèles (figure 5a).

Cependant, à cette étape et selon le cas, nous permettons l'ajout de deux autres contraintes au calcul par MODELLER. La première doit permettre de modéliser la protéine dans le contexte d'un complexe protéique homo ou hétéro oligomérique (figure 5a). Ici, la structure support monomérique est remplacée par sa structure quaternaire probable issue de la base PQS (Protein Quaternary State (Henrick and Thornton, 1998)). Les résidus hydrophobes présents à l'interface sont alors correctement enfouis et cela peut se vérifier lors de l'évaluation des modèles (Douguet, et al., 2002).

La seconde option permet de prendre en compte d'éventuels ligands co-cristallisés avec la protéine support comme un ion ou une petite molécule organique (figure 5b). Les ligands ainsi transférés vont être utilisés comme des contraintes stériques (et uniquement stériques) pour positionner les chaînes latérales des résidus avoisinants. Dans tous les cas, lorsque le modèle sert au criblage de ligands, il sera nécessaire de valider le site actif par l'amarrage de ligands déjà connus.

Alternativement, il est possible de réduire la modélisation à celle du cœur commun. Le programme SCWRL (Dunbrack, 1999) construit un modèle en maintenant fixe le squelette de la protéine support et en transposant tels quels les résidus conservés (y compris la chaîne latérale). Ensuite, il optimise la position de la chaîne latérale des résidus mutés en fonction de leur environnement. Le modèle produit est partiel puisqu'il manque les zones d'insertion/délétion mais il peut être suffisant pour l'analyse du site actif qui est souvent le cœur conservé de la protéine.

La figure 5c montre la page HTML des résultats de la modélisation de la phospholipase $A_2$ sécrétée humaine de groupe III. Trois modèles ont été générés par MODELLER. Chaque modèle est évalué par la fonction objective de MODELLER



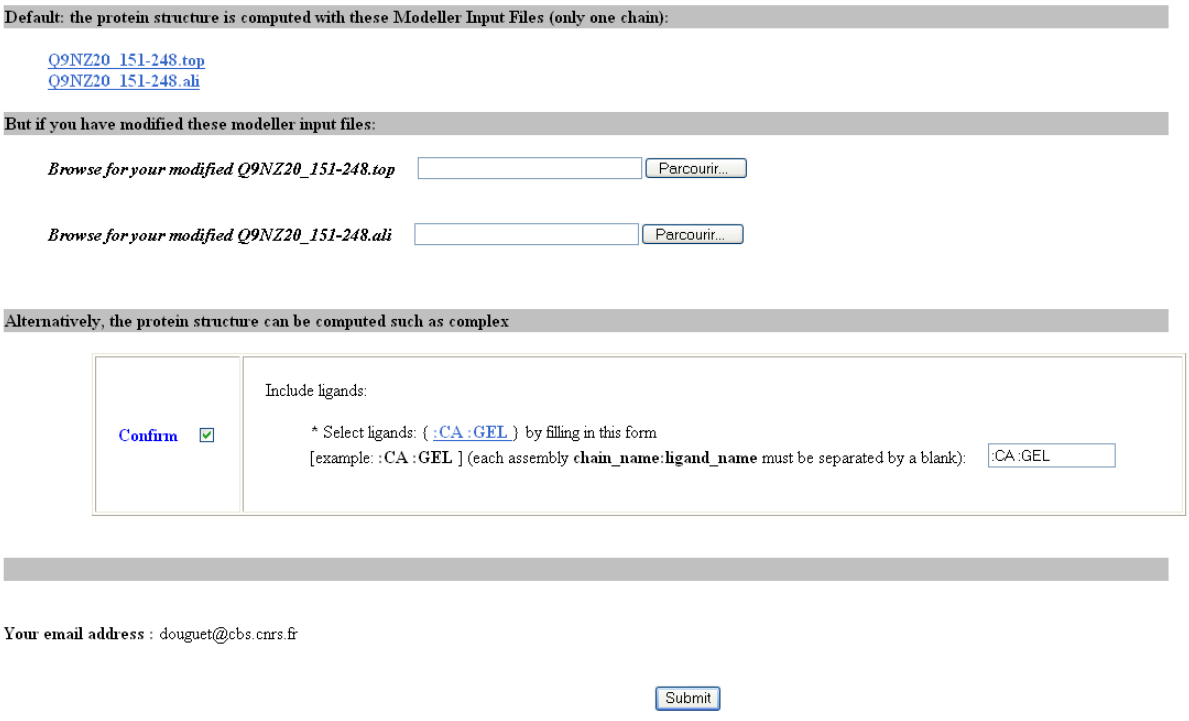

Total_Fragments ( 98) : PKB = -34818 (-355.29 /residue)

**Figure 4**. Etape 2°. Evaluation de la compatibilité 1D/3D entre la séquence soumise (Q9NZ20_151-248) et la structure support 1POC. Chaque fragment est noté par TITO. Les fragments sont séparés par des insertions/délétions. Ici, tous les fragments ont un score négatif favorable. Le score total est de -34818 pour 98 résidus alignés. L'identité de séquence est de 44%. Les résidus conservés sont mis en gras et ceux mutés sont colorés selon leur classe physico-chimique : les hydrophobes en jaune, les polaires en vert, les chargés positivement en bleu, les chargés négativement en rouge et les aromatiques en marron. Les résidus catalytiques sont conservés et identifiés par une étoile rouge.

**Figure 5**. Etape 3°. Préparation des fichiers d'entrée pour la modélisation par homologie. a) l'utilisateur peut choisir entre la modélisation du cœur commun par SCWRL ou bien la modélisation complète par MODELLER. Dans ce dernier cas, il est possible de préférer une modélisation dans un contexte quaternaire en utilisant la protéine support issue de la MSD/PQS. b) Afin de conserver l'orientation des chaînes latérales du site actif, il est préférable d'inclure les ligands de la structure support. Ici, le calcium CA et le 1-o-octyl-2-heptylphosphonyl-*sn*-glycero-3- phosphoethanolamine GEL serviront de contraintes stériques. c) Page de résultats de la modélisation par MODELLER. Chaque modèle est associé à l'énergie de la fonction objective MODELLER (plus elle est faible, meilleure est l'adaptation aux contraintes) et au score moyen d'évaluation par VERIFY-3D et PROSAII.

a)


b)



c)

**Download Models for Q9NZ20_151-248 (PDB Files)**

**Models :**

Model 1 (to edit file) **Energy = 898.3123**

See the template multi-evaluation **pdb1poc.ent**
See the model1 multi-evaluation **model1_Q9NZ2.pdb**

| PDB file | Chain | PROSAII | VERIFY3D |
|---|---|---|---|
| pdb1poc.ent | - | -0.638 (85 aa) | 0.381 (134 aa) (Verify3D server) |
| model1_Q9NZ2.pdb | 0 | -0.032 (49 aa) | 0.251 (98 aa) (Verify3D server) |

Model 2 (to edit file) **Energy = 1307.3424**

See the template multi-evaluation **pdb1poc.ent**
See the model2 multi-evaluation **model2_Q9NZ2.pdb**

| PDB file | Chain | PROSAII | VERIFY3D |
|---|---|---|---|
| pdb1poc.ent | - | -0.638 (85 aa) | 0.381 (134 aa) (Verify3D server) |
| model2_Q9NZ2.pdb | 0 | 0.091 (49 aa) | 0.263 (98 aa) (Verify3D server) |

Model 3 (to edit file) **Energy = 1421.3796**

See the template multi-evaluation **pdb1poc.ent**
See the model3 multi-evaluation **model3_Q9NZ2.pdb**

| PDB file | Chain | PROSAII | VERIFY3D |
|---|---|---|---|
| pdb1poc.ent | - | -0.638 (85 aa) | 0.381 (134 aa) (Verify3D server) |
| model3_Q9NZ2.pdb | 0 | 0.030 (49 aa) | 0.188 (98 aa) (Verify3D server) |

**The lower the energy, the better the model**

**Etape 4 : Evaluation du modèle**

La validation la plus classique consiste à inspecter la stéréochimie du modèle 3D, c'est-à-dire de vérifier qu'il n'y a pas de violations des règles de base observées sur les protéines globulaires. La projection de Ramachandran est la plus connue. Elle représente la distribution des angles Phi et Psi des acides aminés composants la protéine.

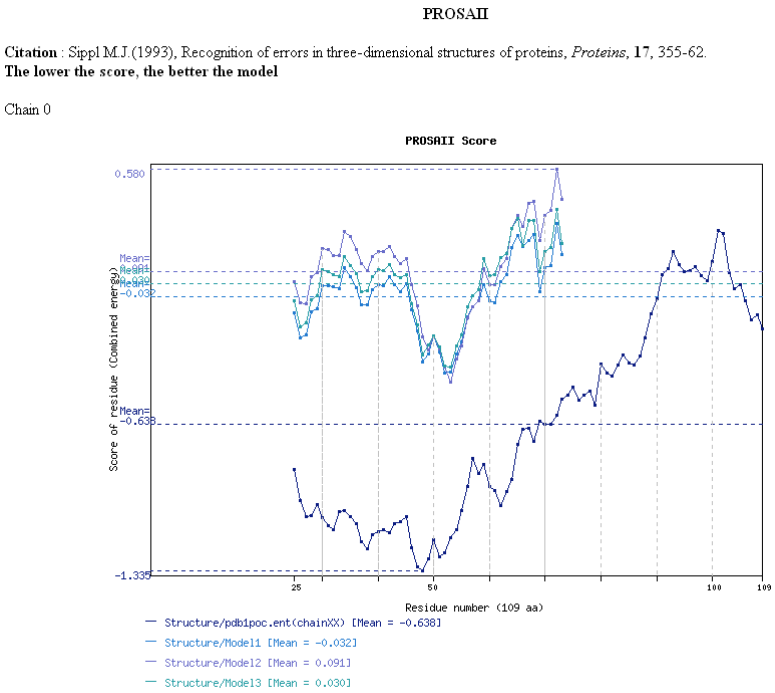
D'autres méthodes d'évaluation calculent la probabilité d'un résidu à se trouver dans un environnement donné. Le serveur @TOME utilise plus particulièrement les programmes PROSA-II (Hendlich, et al., 1990) et Verify-3D (Bowie, et al., 1991). Une partie de ce travail a été réalisée par Emma Ribes lors de son stage maîtrise. Ces méthodes analysent et comparent les distributions statistiques des carbones α et des résidus, respectivement, pour repérer les segments mal définis. De nombreuses autres méthodes sont maintenant accessibles en ligne :

- ERRAT (http://www.doe-mbi.ucla.edu/Services/ERRAT/ ; (Colovos and Yeates, 1993))
- MolProbity (http://molprobity.biochem.duke.edu/ ; (Davis, et al., 2004))
- WHAT IF (http://swift.cmbi.kun.nl/WIWWWI// ; (Rodriguez, et al., 1998)),
- Eval123D (http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_valid.html ; SFE (Chiche, et al., 1990), EvTree (Gelly, et al., 2005), Eval23D (Gracy, et al., 1993)).
- ProQ (http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi ; (Wallner and Elofsson, 2003))

Hormis la valeur du score de chaque modèle, la comparaison du score de la structure support à celui du modèle permet d'évaluer la qualité globale de ces derniers mais aussi de détecter les segments moins bien compatibles (figures 6a et 6b).

**Figure 6**. Etape 4°. Evaluation des modèles provenant de MODELLER. Chaque modèle est automatiquement évalué par VERIFY-3D et PROSAII. Le score global ('Mean') est comparé au score de la protéine support 1POC (voir figure 5c). Deux graphes complètent l'évaluation en représentant le score pour chaque résidu. Les trois modèles et la structure support sont représentés dans le même référentiel afin de faciliter l'analyse. Le modèle 1 de meilleur énergie MODELLER est aussi celui qui a le meilleur score PROSAII et celui pour lequel les 70 premiers résidus (sur 98) sont les mieux évalués par VERIFY-3D (bien que le score global est meilleur pour le modèle 2 (0.263 versus 0.251)). Le score moyen VERIFY-3D est acceptable pour un modèle (entre 0.3-0.4 pour de bons modèles) mais celui de PROSAII est faible (entre -0.8 et -1 pour de bons modèles). Cependant, cette dernière remarque peut être apportée à l'évaluation de la structure expérimentale elle-même. Son score PROSAII est médiocre pour une structure expérimentale (-0.638). Elle est due à la partie C-ter et plus particulièrement aux 35 derniers résidus de la partie C-ter. Cette partie peu structurée est une boucle adoptant une pseudo-structuration en feuillet β (maintenue par un pont disulfure) n'interagissant avec le reste de la protéine que par les 10 derniers résidus pour former un feuillet β. Les modèles ne sont que très légèrement améliorés, selon PROSAII, lorsque l'on modélise un C-ter plus long. Cependant, le site actif est lui plus en amont en séquence et n'est pas directement perturbé par cette partie du C-ter. Enfin, la modélisation de hGIII dans le contexte homodimérique proposé par la PQS n'a pas d'effet sur la structuration du site actif.

a)

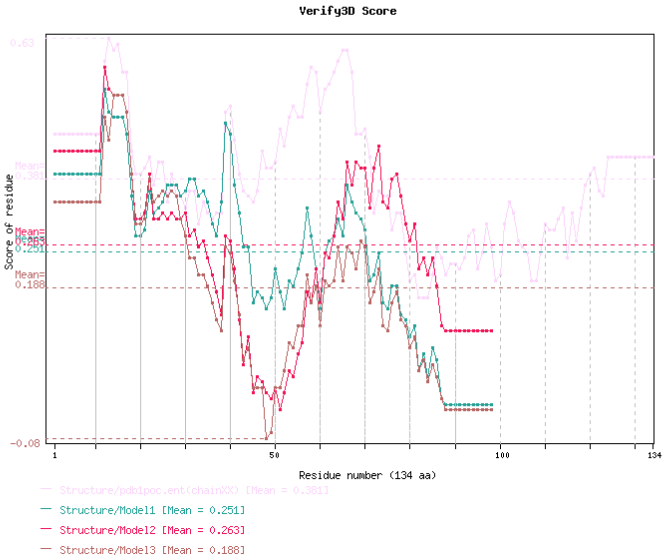


b)

**VERIFY3D**

**Citation** : Eisenberg D, Luthy R, Bowie JU. (1997), VERIFY3D: assessment of protein models with three-dimensional profiles, *Methods Enzymol.*, **277**, 396-404.
**The higher the score, the better the model**

Chain 0



## Exemple d'application au sein du laboratoire

Courant 2002, le méta-serveur fut plus particulièrement utilisé à l'étude d'une Polyphosphate/ATP-NAD kinase de *Mycobacterium tuberculosis* dans le cadre d'une collaboration, à l'initiative de Gilles Labesse, avec L. Assairi et A.-M. Gilles de l'Institut Pasteur. Cette protéine est une cible potentielle pour de nouveaux anti-tuberculeux puisqu'elle serait la seule enzyme pouvant catalyser le transfert d'un phosphate au NAD (Nicotinamide Adénine Dinucléotide) et qu'elle possède deux poches accommodants le NAD et l'ATP. Jusqu'alors, aucune donnée structurale n'existait à part une faible similarité de séquence avec des PFKs (6-phosphofructokinases) dont la structure expérimentale était accessible (code pdb 1PFK). Une similarité structurale pouvait alors être envisagée. Au final, il fut découvert une nouvelle super-famille structurale portant une signature originale pour le site de fixation de l'ATP. Cette étude a montré que les diacylglycéride kinases (DGKs), les sphingosine kinases (SKs) impliquées dans l'apoptose et les PPN kinases (PPNK) adoptent le même repliement que la 6-phosphofructokinase malgré un faible taux d'identité de séquence allant de 10 à 20% sur 250 acides aminés (Labesse, et al., 2002). L'utilisation du méta-serveur permit de confirmer les résultats de l'analyse de séquence en évaluant rapidement la compatibilité du repliement de la PFK avec des séquences de DGK, SK et PPNK. En complément, la recherche de motifs par PHI-BLAST (Zhang, et al., 1998) et PATTINPROT (Combet, et al., 2000) permis d'affiner la signature commune du site de fixation de l'ATP. Un certains nombre de résultats issus d'expérience de mutagenèse dirigée soutenaient ces hypothèses. Depuis, la structure de NADK de *Mycobacterium tuberculosis* a été déterminée expérimentalement (code PDB 1U0R, 1U0T, 1Y3I (co-cristallisée avec le NAD) et 1Y3H) et son étude se poursuit au laboratoire dans le but notamment d'identifier des inhibiteurs.

## Evaluation du serveur @TOME

Les sessions CASP (Critical Assessment of Protein Structure Prediction) sont des sessions internationales d'évaluation des méthodes de prédiction de la structure des protéines (http://predictioncenter.llnl.gov/). Elles donnent la séquence des protéines mises en jeu et dont la structure expérimentale vient d'être déterminée mais retenue jusqu'à l'expiration du 'concours'. Plusieurs modèles peuvent être soumis pour chaque cible. Ces sessions ont débuté en 1994 (CASP1). En treize années, les CASP ont permis de quantifier les progrès et de faire, tous les deux ans, un point sur l'état de l'art. Parallèlement, depuis 1998, il existe des sessions CAFASP dont l'accès est restreint aux prédictions purement automatiques par des serveurs. On distingue 3 catégories de prédictions : la modélisation comparative (CM, i.d.~30%), la reconnaissance de repliements homologues (FR(H) ; protéines ayant un ancêtre commun), la reconnaissance de repliements analogues (FR(A) ; protéines sans filiation commune) et les nouveaux repliements (NF ou *ab initio*), catégorie la plus difficile puisque la notion de support expérimental disparaît.

Depuis la première session CASP, il a été noté de nets progrès, bien que lents, au niveau de la qualité de l'alignement des séquences, de la reconnaissance de repliement (FR) et de la modélisation des protéines de repliement nouveau (NF). L'alignement des séquences reste la principale source d'erreur influençant sur la qualité des modèles surtout en deçà d'un taux d'identité de séquence de 30%. Au-delà de 60%, l'erreur globale est inférieure à 1 Å pour les carbones $\alpha$. Ainsi, une erreur sur l'alignement d'un acide aminé dans une séquence peut engendrer une erreur de 3.8 Å sur les carbones $\alpha$ (et par conséquent, une erreur sur la position de 4 acides aminés peut provoquer un RMSD > 12 Å ; (Venclovas, et al., 2001)).

La session du CASP5 eu lieu durant l'été 2002 pour 67 cibles (de T0129 à T0195 ; http://predictioncenter.gc.ucdavis.edu/casp5). Il était important d'y participer afin d'évaluer les performances et les limites du serveur @TOME surtout lorsqu'il est utilisé en mode non-expert (en automatique) pour, spécifiquement, des cibles de la catégorie CM. Le second groupe, TOME (G. Labesse) représentait le mode semi-automatique (expert) du serveur en couplage avec d'autres outils comme VITO (Catherinot and Labesse, 2004).

Globalement, @TOME et TOME se sont bien défendu ce qui nous valu une invitation à la conférence CASP5 à Asilomar. En effet, ils ont été classés 9$^{ième}$ et 26$^{ième}$ (pour TOME et @TOME, respectivement) parmi les 40 premiers groupes sur 187 inscrits (dont 72 serveurs au CAFASP3 (Moult, et al., 2003)). Ci-dessous, la Table 1 indique les classements basés sur un Z-score calculé par l'équipe de M. Levitt sur la totalité des modèles soumis (http://www.forcasp.org/print/1983/). le Z-score est une mesure de la déviation par rapport à la moyenne (un Z-score de 0 correspond donc au cas moyen ; plus il est élevé plus la performance est remarquable).

La première conclusion est que quelque soit la catégorie, de 1 à 3 groupes d'excellence se démarquent du reste. David Baker est spécialiste de l'*ab initio* (NF) et le groupe des 'polonais' se fait remarquer pour la modélisation comparative et la reconnaissance de repliement (Ginalski, Bujnicki et leur serveur bioinfo.pl). Sur l'ensemble des cibles, @TOME est bien classé mais lui et les autres serveurs sont nettement supplantés par les artisans. Ce résultat était attendu mais il permet à chaque CASP de mesurer l'écart entre ce qui peut être réalisé par un expert et un non-expert. La catégorie CM à haute identité de séquence relève de ce cas de figure comme nous allons le montrer à partir des résultats d'@TOME.

La table 2 résume les résultats pour les 28 cibles de la catégorie CM et CM/FR(H) composées d'un seul domaine. Les premières colonnes contiennent le nom de chaque cible, la longueur de la séquence, l'espèce, la méthode utilisée pour résoudre la structure, le code PDB s'il existait déjà, la composition en domaine, la catégorie et une brève description du repliement/famille de la protéine. Les quatre colonnes suivantes indiquent les résultats du serveur @TOME en mode automatique complet. Ce mode automatique sélectionnait jusqu'à 10 supports distincts pour construire 40 modèles à partir du meilleur alignement (10 * 3 modèles par MODELLER + 10 modèles du cœur commun par SCWRL). Les 5 meilleurs modèles après validation par PROSAII et VERIFY-3D étaient soumis au CASP. A noter, que la fonction de sélection, empirique, n'avait auparavant jamais été testée.

La mesure utilisée pour classer les modèles est le GDT-TS (Moult, et al., 2001). De nombreuses superpositions, entre le modèle et la structure expérimentale, sont testées et celle qui produit le plus grand nombre de résidus superposés (C$\alpha$) en deçà du seuil choisi (1, 2, 4 et 8 Å) est utilisée pour le calcul final. Dans le cadre de la catégorie CM, on s'attend à ce que la plupart des résidus appartiennent au seuil 8 et un peu moins au seuil 4. Les seuils 1 et 2 seront, eux, les témoins des variations locales du modèle. Le GDT-TS, bien qu'imparfait, est plus adapté que le RMS ('Root Mean Square') classique pour évaluer et capturer la qualité globale d'un modèle. En effet, un modèle complètement faux peut, par chance, obtenir un bon RMS et inversement un médiocre RMS peut correspondre à un bon modèle. Un exemple de ce type est donné ci-après pour la cible T0183 modélisée par @TOME. Le GDT-TS est une mesure présentée en pourcentage.

GDT-TS = 1/4 [ N1 + N2 + N4 + N8]

Nn est le nombre de résidus superposés en deçà du seuil n

***Succès :***

- 10 cibles ont une identité de séquence supérieure à 30% avec la structure support expérimentale (lignes grisées dans la Table 2). Toutes ont un GDT-TS supérieur à 73% et la plupart un bon RMS-CA, exceptée T0183 (RMS-CA=18.27 Å), T0151 (RMS-CA=5.36 Å) et T0167 (RMS-CA=4.10 Å). Le cas de T0183 illustre très bien le défaut du RMSD séquence dépendante. La figure 7 présente la superposition de la structure expérimentale de T0183 (PDB1o0y) avec le meilleur modèle soumis par @TOME. La superposition est satisfaisante comme l'indique le score GDT-TS (76.21%). Une superposition basée sur le RMS-CA ne produit pas le même résultat (plus de 18 Å) car l'optimisation de cette mesure passe par une superposition des C-ter afin de minimiser la distance, et cela, au dépend d'une superposition quasi-parfaite du reste de la protéine.

  Le succès d'@TOME en mode automatique dans le contexte de la modélisation d'un seul domaine avec un support homologue à 30% minimum, confirme qu'un non-expert peut produire des modèles de qualité satisfaisante par l'intermédiaire d'un serveur tel que le nôtre.

- 4 cibles : T0142 (26%), T0160 (22%), T0176 (26%) T0178 (27%) sont parmi les cibles pour lesquelles l'identité de séquence avec la structure support est > 20% et pour lesquelles les modèles soumis par @TOME sont satisfaisants (utilisation du support attendu, GDT-TS de 71%, 80%, 51% et 78% et RMS-CA de 3.4 Å, 2.59 Å, 5.91 Å et 3.1 Å). On note cependant que le modèle de T0176 (GDT-TS de 51%) est de qualité moindre malgré qu'il soit classé 17$^{ième}$ sur les 148 soumis. Le meilleur modèle possède un GDT-TS de 54% et un RMS-CA de 4.94 Å. Cela s'explique par une différence entre la topologie de la structure support (PDB1jrm) et celle de la structure expérimentale de T0176 qui ne pouvait être correctement prédite. Une boucle pincée au cœur de la protéine dans 1jrm est extériorisée dans T0176 (PDB1O0y).

***Succès partiels :***

- La cible T0140 est un cas particulier puisqu'il s'agit d'une chimère synthétique dont les extrémités C-terminale et N-terminale devaient être construits par modélisation comparative en utilisant deux supports différents. @TOME sélectionna le support 1mjc (57% d'identité de séquence) pour modéliser le N-ter comme cela était attendu. La modélisation du C-ter aurait nécessité un autre support. Cette modélisation multi-support n'est pas disponible en mode automatique. Ici, aucun des assesseurs n'a fourni de modèles satisfaisants.

- 8 cibles : T0132 (16%), T0133 (13%), T0138 (19%), T0152 (15%), T0157 (15%), T0169 (9%), T0189 (14%), T0192 (16%) %) font partie de la catégorie FR avec un taux

d'identité de séquence de 9 à 20% avec la protéine support la plus adaptée. On distingue 2 groupes : ceux pour lesquels la protéine support a été correctement identifiée (T0132, T0133, T0157 et T0189) et ceux pour lesquels la protéine support utilisée est de la même famille que celle attendue mais dont le domaine/fonction est différent (T0138, T0152, T069 et T0192).

Le GDT-TS varie, dans ce groupe de 9, entre 47% et 60%. Les modèles d'@TOME sont cependant assez bien classés (par exemple, T0133 est classé 2$^{ieme}$ sur les 386 modèles soumis ; figure 8 et 9).

Ces résultats font apparaître un certain succès d'@TOME dans la catégorie FR. Ils montrent cependant qu'en deçà de 30% d'identité de séquence, et surtout en deçà de 20%, la modélisation nécessite une intervention sur le choix du ou des supports en fonction, notamment, de la connaissance de la fonction de la protéine. La modification de la topologie telle qu'elle a été vue pour T0176 n'était pas non plus prévisible (26% d'identité de séquence avec PDB1jrm).

***Echecs :***

- 5 échecs (T0130, T0141, T0156, T0165 et T0195) ont pour origine une erreur dans le choix du support. Les protéines supports avaient un taux d'identité de séquence de 18%, 26%, 14%, 18% et 18%, respectivement. Dans ces cas là, le GDT-TS de nos modèles ne dépasse pas 35% (sauf 44% pour T065). Pour T0130 et T0165, des erreurs d'alignement et/ou leur évaluation ont écarté les supports 1fa0 et 1a8s. Pour T0141, T0156 et T0195, le support attendu (1lba, 1dik et 1jjf, respectivement) a bien été utilisé pour la modélisation par MODELLER mais l'évaluation des modèles résultants ne les a pas classés parmi les 5 premiers. La cible T0141, malgré ces 26% d'identité de séquence avec PDB1lba, présente des changements structuraux au niveau du site actif. Cela explique que même le meilleur modèle (sur 145) n'a qu'un GDT-TS de 45.72%.

| Catégorie (nombre de cibles) | @TOME | | TOME | | Meilleur groupe | |
|---|---|---|---|---|---|---|
| | Rang | Z-score | Rang | Z-score | Rang=1 | Z-score |
| CM* (42) | 28 | 15.72 | 16 | 21.98 | Bujnicki-Janusz | 47.17 |
| FR* (13) | 27 | 5.60 | 30 | 4.82 | Ginalski/bioinfo.pl | 24.26 |
| NF* (13) | 38 | 3.18 | 9 | 7.64 | David Baker | 25.72 |
| **Global** (68) | **26** | **25.09** | **9** | **37.35** | Ginalski/bioinfo.pl | **75.94** |

**Table 1**. Le rang et le Z-score d'@TOME, TOME et du premier groupe classé sont indiqués pour chaque catégorie par ordre croissant de difficulté (CM < FR < NF).
* Seul le premier modèle est pris en compte. Dans le cas d'@TOME, le premier modèle fut rarement le meilleur ce qui lui accorde un Z-score assez faible pour CM, FR et NF mais un assez bon Z-score pour le Global qui prend en compte les 5 modèles soumis. D'autre part, ces résultats prennent en compte les cibles pour lesquelles il y a plus d'un domaine. @TOME est dévaforisé dans ce cas puisqu'il n'effectuait pas de découpage en domaines.

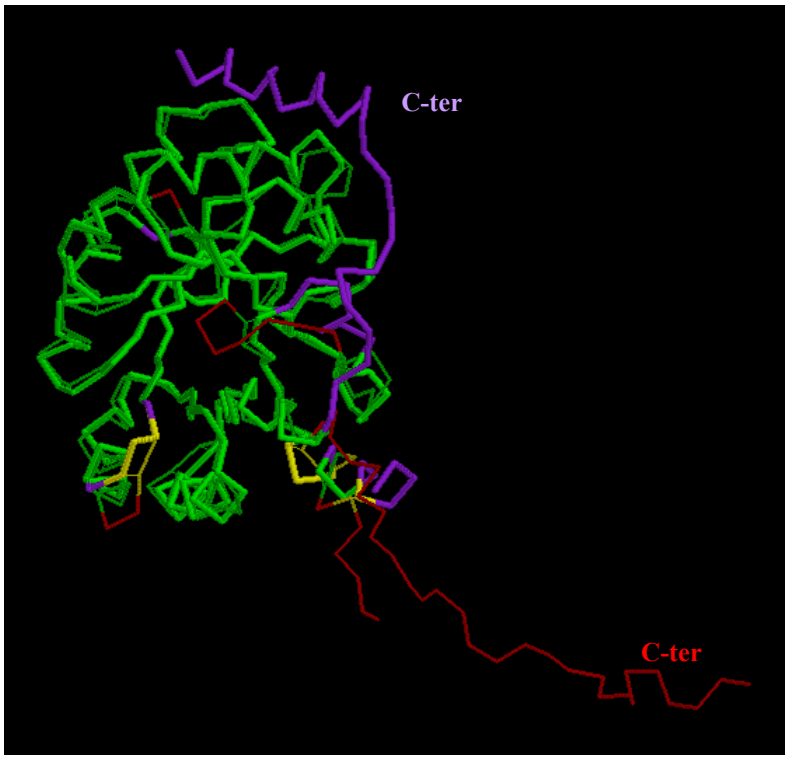



**Figure 7**. Superposition de la structure expérimentale de T0183 avec le meilleur modèle d'@TOME. Les segments verts sont correctement alignés, les segments en jaunes sont plus éloignés mais à moins de 4 Å de distance, les segments rouges non alignés appartiennent au modèle et les segments mauves non alignés appartiennent à la structure expérimentale. La superposition est satisfaisante comme l'indique le score GDT-TS (76.21%). Une superposition basée sur le RMS-CA ne produit pas le même résultat car l'optimisation de cette mesure passe par une superposition des C-ter afin de minimiser la distance, et cela, au dépend d'une superposition quasi-parfaite du reste de la protéine. Le cas présent illustre bien l'avantage du GDT-TS .

**Table 2**. Liste des cibles CM et CM/FR mises en jeu lors du CASP5. Les premières colonnes contiennent le nom de chaque cible, la longueur de la séquence, l'espèce, la méthode utilisée pour résoudre la structure, le code PDB, la composition en domaine, la catégorie et une brève description repliement/famille de la protéine. Les quatre colonnes suivantes indiquent les résultats du serveur @TOME (groupe 464) : le score GDT-TS, le RMS-CA, le classement du meilleur modèle sur les 5 soumis par @TOME et une remarque concernant la réussite ou l'échec de l'identification de la structure support appropriée.

| Target ID | Length | Name Species | Method | PDB ID | Domains Range | CASP Class | Description | ATOME groupe 464 GDT-TS | ATOME groupe 464 RMS-CA | ATOME groupe 464 rank | ATOME groupe 464 remarque |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **T0130** | 114 | HI0073 *H.influenzae* | X-ray | | single | CM/FR(H) | Nucleotidyltransferase superfamily. Sequence finds PDB (**18% 1fa0**, Dali Z-score 5.0) with transitive PSI-BLAST searches. Compared to 1fa0, structure contains generally shorter structural elements and loses a $\beta$-hairpin from one edge while retaining the active site. | 27 | 12.67 | Le meilleur modèle : GDT-TS=59.25 et RMS-CA=6.87 | **ECHEC** erreur support |
| **T0132** | 154 | HI0827 *H.influenzae* | X-ray | | single | CM/FR(H) | Thioesterase superfamily member. Sequence finds 4-Hydroxybenzoyl Coa Thioesterase (**16%, 1bvq**, Dali Z-score 13.4) with transitive PSI-BLAST searches. | 60.2 | 6.39 | **17ième / 398 modèles** | bon support |
| **T0133** | 312 | HIP1R N-terminal domain *rat* | X-ray | | single | CM/FR(H) | $\alpha/\alpha$ superhelix fold, the same family as N-terminal domain of phosphoinositide-binding clathrin adaptor (**13% 1hg5**). | 52.82 | 11 | **2ième / 386 modèles** (TOME est 1er) | bon support |
| **T0137** | 133 | Fatty acid binding protein *E.granulosus* | X-ray | | single | CM | Fatty acid-binding protein family (**43% 2ans**) | 94.36 | 1.19 | | bon support homologue 1lic (43.5%) |
| **T0138** | 135 | KaiA N-terminal domain *S.elongatus* | NMR | 1m2e 1m2f | single | FR(H) | Flavodoxin-like fold, CheY-like superfamily (**19% 1kgs**). Homology inference based on structural similarity (Dali Z-score 13.2). | 54.08 | 7.42 | **Modèles bien classés** (le meilleur : GDT-TS=67.59 et RMS-CA=4.83) | support homologue sauf domaine 3chy (11%) |
| **T0140** | 103 | 1B11 synthetic protein | X-ray | 2bh8 | single, composite of 2 chains B18-B74, A75-A102 | CM | **Synthetic protein** composed of cold shock protein A (N-terminal) and E.coli 30S ribosomal subunit protein S1 (C-terminal). While each parent structure forms an OB fold, the synthetic protein forms an OB-fold-like | 47.67 | 14.75 | **24ième / 424 modèles** | bon support partiel 1mjc (57%) |
| **T0141** | 187 | AmpD *C.freundii* | NMR | 1iya | single | CM/FR(H) | Homologue of T7 lysozyme (**26% 1lba**). Unexpected structural differences in the active site region | 16.18 | 23.44 | Le meilleur modèle : GDT-TS=45.72 et RMS-CA=7.59 | **ECHEC** erreur support |
| **T0142** | 282 | Nitrophorin *C.lectularius* | X-ray | | single | CM | DNaseI-like fold, Inositol polyphosphate 5-phosphatase family (**26% 1i9z**). | 71.34 | 3.4 | **15ième / 371 modèles** | bon support |
| **T0150** | 102 | Ribosomal protein L30E *T.celer* | X-ray | 1h7m | single | CM | L30e/L7ae ribosomal protein family (**34% 1ck2**) | 73.96 | 2.37 | | bon support homologue 1cn7 (32.3%) |
| **T0151** | 164 | Single-strand binding protein *M.tuberculosis* | X-ray | | single | CM | ssDNA-binding protein, OB-fold (**30% 1qvc**) | 75.71 | 5.36 | **13ième / 379 modèles** | bon support homologue 1kaw (32.3%) |
| **T0152** | 210 | Hypothetical protein Rv1347c *M.tuberculosis* | X-ray | | single | CM/FR(H) | Acyl-CoA N-acyltransferase (NAT) family (**15% 1i1d**). Typical for this family $\beta$-bulge in the active site lacks in this protein leading to a different shape of the $\beta$-sheet. | 47.47 | 24.89 | **8ième / 413 modèles** | support homologue sauf domaine 1cjw (16%) |
| **T0153** | 154 | dUTPase *M.tuberculosis* | X-ray | 1mq7 | single | CM | dUTPase, beta-clip fold (**35% 1eu5**) | 87.69 | 1.61 | **4ième / 371 modèles** | bon support homologue 1euw (34%) |
| **T0155** | 133 | Probable dihydroneopterin aldolase *M.tuberculosis* | X-ray | | single | CM | 7,8-dihidroneopterin aldolase, T-fold (**33% 1dhn**) | 96.80 | 0.91 | **3ième / 344 modèles** | bon support homologue 2dhn (33.3%) |
| **T0156** | 157 | Probable SAM-dependent methyltransferase *M.tuberculosis* | X-ray | | single | FR(H) | Phosphohistidine domain superfamily, the "swiveling" domain fold (**14% 1dik**, Dali Z-score 7.5). Homology inferred from structural similarity. | 18.43 | 15.54 | Le meilleur modèle : GDT-TS=37.66 et RMS-CA=9.45 | **ECHEC** erreur support |
| **T0157** | 138 | yqgF *E.coli* | X-ray | | single | FR(H) | Ribonuclease H-like superfamily (**15% 1hjr** Dali Z-score 9.6). Homology inferred from the presence of described RNaseH motifs. | 55.84 | 5.26 | **27ième / 426 modèles** | bon support |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **T0160** | 128 | VAP-A protein *rat* | X-ray | | single | CM | Major sperm protein family, Immunoglobulin-like fold (**22% 2msp**) | 80.2 | 2.59 | | bon support homologue 1msp (19%) |
| **T0165** | 318 | Cephalosporin C deacetylase *B.subtilis* | X-ray | | single | CM/FR(H) | $\alpha/\beta$-hydrolase superfamily (**18% 1a8s**) | 44.81 | 12.18 | **Modèles bien classés** (le meilleur : GDT-TS=53.22 et RMS-CA=12.22) | erreur support (mais fold et superfamille correct) |
| **T0167** | 185 | Hypothetical cytosolic protein yckF | X-ray | | single | CM | SIS domain (**39% 1jeo**) | 77.36 | 4.10 | | bon support |
| **T0169** | 156 | yqjY *B.subtilis* | X-ray | | single | CM/FR(H) | N-acetyl transferase (NAT) family (**9% 1bo4**) | 60.42 | 4.23 | **Modèles bien classés** (le meilleur : GDT-TS=70.67 et RMS-CA=5.41) | support homologue sauf domaine 1b87 (19%) |
| **T0176** | 100 | Hypothetical protein yggU *E.coli* | X-ray | | | CM | Close homologue of hypothetical protein MTH637 (**26% 1jrm**) | 51.50 | 5.91 | **17$^{ième}$ / 148 modèles** (le meilleur : GDT-TS=54 et RMS-CA=4.94) | bon support |
| **T0178** | 219 | phosphate aldolase *A.aeolicus* | X-ray | | single | CM | Deoxyribose-phosphate aldolase DeoC, TIM-barrel (**27% 1jcl**) | 78.31 | 3.1 | | bon support homologue 1ktn (28.5%) |
| **T0182** | 250 | TM1478 *T.maritima* | X-ray | 1o0x | single | CM | Methionine aminopeptidase (**42% 2mat**) | 92.27 | 1.32 | | bon support |
| **T0183** | 248 | TM1559 *T.maritima* | X-ray | 1o0y | single | CM | Deoxyribose-phosphate aldolase DeoC, TIM-barrel (**30% 1jcl**) | 76.21 | 18.27 | | bon support homologue 1ktn (30.2%) |
| **T0188** | 124 | TM1816 *T.maritima* | X-ray | | single | CM | Close homologue of hypothetical protein MTH1175, RNaseH-like fold (**31% 1eo1**) | 75 | 2.56 | | bon support |
| **T0189** | 319 | TM0828 *T.maritima* | X-ray | | single | CM/FR(H) | Ribokinase-like family (**14% 1rk2**) | 53.69 | 5.61 | **26$^{ième}$ / 367 modèles** | bon support homologue 1rkd (16%) |
| **T0190** | 114 | Transthyretin-related protein *E.coli* | X-ray | | single | CM | Prokaryotic homologue of transthyretin (**31% 1dvx**) | 88.74 | 1.92 | | bon support homologue 1etb (16%) |
| **T0192** | 171 | Spermidime/Spermine acetyltransferase *human* | X-ray | | single, composite of 2 chains 2-153 (first chain), 154-171 (second chain) | CM | N-acetyl transferase (NAT) family, (**16% 1qsm**), domain-swapped last strands | 47.5 | 9.44 | Modèles moyens (le meilleur : GDT-TS=69.12 et RMS-CA=3.18) | support homologue sauf domaine 1b87 (12%) |
| **T0195** | 299 | Hypothetical esterase in SMC3-MRPL8 intergenic region *S.cerevisiae* | X-ray | | single | CM/FR(H) | $\alpha/\beta$-Hydrolase superfamily (**18% 1jjf**) | 35.86 | 7.93 | Modèles moyens (le meilleur : GDT-TS=62.67 et RMS-CA=5.85) | **ECHEC** erreur support (mais fold et superfamille correcte) |

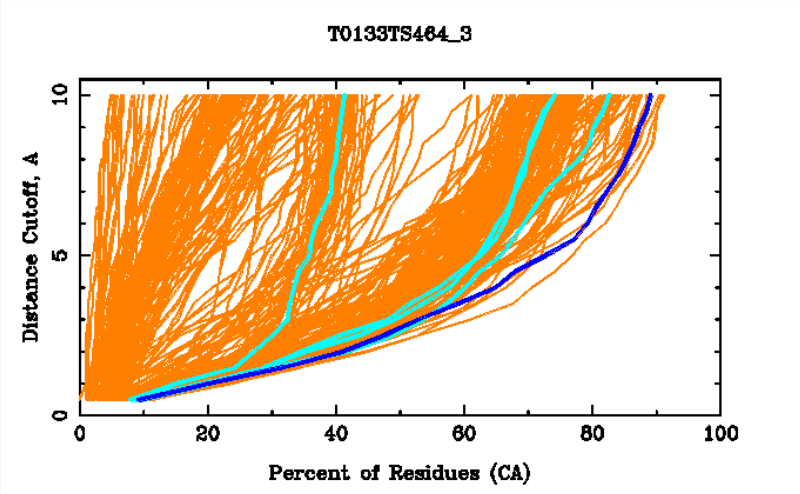
**3D Structures - Sequence Dependent Analysis**
**All 3D models for target: T0133**

**Group name: 464**

**Total number of ALL models submitted for target T0133 : 386**
**Number of models submitted by group 464 : 5**

BLUE - Model - T0133TS464_3
CYAN - **Other models submitted by group: 464**
BROWN - **Models submitted by other groups**

GDT analysis: largest set of CA atoms (percent of the modeled structure) that can fit under DISTANCE cutoff 0.5A, 1.0A, 1.5A, ... , 10.0A




Largest set of residues from the Model T0133TS464_3 that can fit under **DISTANCE cutoff** :
20 percent of residues ( 59 residues ) fits under 1A with local RMSD 0.67
41 percent of residues ( 120 residues ) fits under 2A with local RMSD 1.27
65 percent of residues ( 190 residues ) fits under 4A with local RMSD 2.24
79 percent of residues ( 232 residues ) fits under 6A with local RMSD 3.15
Percent of the structure predicted: 100.00 ( CA atoms: Model 293 , Target 293 )

**Figure 8**. Exemple de résultat fournit par les assesseurs de CASP5. Le graphe permet de comparer le GDT de tous les modèles soumis pour la cible T0133 (386 au total). Les lignes colorées correspondent aux 5 modèles soumis par @TOME. La ligne bleu foncé correspond au meilleur modèle classé ici 2$^{ième}$. Le meilleur modèle fut soumis par TOME (G. Labesse, mode expert). Le meilleur GDT-TS correspond à la ligne formant la plus grande aire.

## Conclusions du CASP5 et état de l'art en 2007 (CASP6 et CASP7)

Si la modélisation artisanale ne progresse presque plus, la modélisation toute automatique dans la catégorie CM a quant à elle significativement évoluée (Fischer, et al., 2003; Kryshtafovych, et al., 2005). Dans cette catégorie, les serveurs sont capables de fournir des modèles de qualité 'artisanale'. Les sessions CASP4 (2000) et CASP5 (2002) furent marquées par l'apparition de plusieurs serveurs capables d'une prédiction automatique de la reconnaissance de repliement mais aussi de la structure tridimensionnelle des protéines (SDSC par Bourne et Shindyalov lors du CASP4). Plus particulièrement, les méta-serveurs obtinrent de bonnes performances lors du CASP5 (analyse/consensus des résultats de serveurs indépendants dans le domaine de la reconnaissance de repliement). Bien que pouvant être considérés comme des exploiteurs d'autres serveurs, ces méta-serveurs permettent de réaliser un consensus qui retient les principales informations de chaque serveur et pas uniquement le résultat du rang 1. Leur succès ne concerne cependant que la catégorie la plus facile (CM). Les résultats d'@TOME présentés précédemment illustrent et valident notre approche pour cette catégorie de prédiction. Bien entendu, de nombreuses améliorations restent à apporter à notre serveur qui n'a pas évolué depuis 2002 par manque de moyens humains (découpage en domaines, optimisation de l'alignement, modélisation multi-support en automatique, mise à jour du méta-serveur, amélioration dans la modélisation des boucles, du placement des chaînes latérales, des contacts (ou 'packing'), discrimination des modèles…). Et d'une façon générale, il reste encore un long chemin à parcourir avant d'obtenir des modèles comparables, en qualité, aux structures expérimentales (GDT-TS > 90% ; (Tress, et al., 2005; Venclovas, et al., 2003).

Les résultats des approches automatiques (non-expertes) sont beaucoup plus nuancés voire aléatoires pour les catégories FR (reconnaisssance de repliement) et NF (nouveaux repliements). Car, globalement, la qualité des modèles reste inversement proportionnelle à la difficulté (taux d'identité de séquence et 'superposabilité' entre la cible et le support le plus proche) même en mode artisanal (Kryshtafovych, et al., 2005). Cependant, les progrès mesurés au cours des CASP concernent ces 2 catégories.

En complément, voici une liste de serveurs qui proposent d'autres procédures automatiques de modélisation par homologie :

- SWISS-MODEL (Schwede, et al., 2003) : http://swissmodel.expasy.org/workspace/index.php?func=modelling_simple1&userid=USERID&token=TOKEN
- ROBETTA (Chivian, et al., 2003) : http://robetta.bakerlab.org/
- GENO3D (Combet, et al., 2002) : http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html
- CPHmodels : http://www.cbs.dtu.dk/services/CPHmodels/
- 3D-JIGSAW (Bates, et al., 2001) : http://www.bmm.icnet.uk/~3djigsaw/
- MODWEB (Pieper, et al., 2006) : http://salilab.org/modweb

## III - Etude des interactions protéine-protéine

De décembre 2003 à octobre 2004, j'ai été mise à disposition du Laboratoire de Bioinformatique Structurale de l'Université de Stony Brook, dirigé par le Pr. Ilya Vakser (actuellement directeur du Centre de Bioinformatique de l'Université de Kansas). Les membres de ce laboratoire sont reconnus pour leurs compétences dans la prédiction des interactions entre protéines notamment par leur programme GRAMM et leur participation aux sessions CAPRI (Critical Assessment of PRediction of Interactions (http://capri.ebi.ac.uk), analogues aux sessions CASPs présentées précédemment).

La plupart des protéines s'assemblent en complexe pour accomplir leur fonction. Une étude récente a, pour la première fois, montré expérimentalement, chez la levure, que 88% des ~2000 protéines ayant été purifiées étaient sous forme de complexe (Gavin, et al., 2006). Cela illustre l'importance des interactions entre protéines dans une cellule. Actuellement, les principes qui régissent ces interactions ne sont que partiellement compris, assez mal décrits et donc moyennement prédits. Les prédictions peuvent cependant fonctionner comme le montre l'exemple de l'interaction entre le domaine TEM-1 de la β-lactamase avec la 'β-lactamase inhibitory protein' BLIP (Strynadka, et al., 1996).

La prédiction des interactions entre protéines nécessite l'utilisation de programmes capables de prédire la configuration des complexes biologiques. Un certain nombre de travaux ont été réalisés à partir de l'étude des complexes cristallographiques présents dans la PDB (Protein Data Bank) (Carugo and Argos, 1997; Dasgupta, et al., 1997; Janin and Rodier, 1995; Jones and Thornton, 1996; Vajda, et al., 2002). Ces complexes observés expérimentalement comportent des complexes réellement biologiques et des complexes artificiels engendrés par les conditions de cristallisation. La distinction des deux familles n'est pas triviale, et pourtant, indispensable à la réalisation de prédictions fines.

La notion de prédiction implique l'analyse de données rassemblées, de préférence, en bases de données de complexes protéiques observés expérimentalement. Ces bases sont essentielles pour étudier les interfaces protéiques et développer des programmes de prédiction (voir (Gray, 2006; Mendez, et al., 2005) pour une revue plus exhaustive des programmes et méthodes actuels). Ces derniers sont la combinaison d'une procédure de recherche et d'une fonction d'évaluation.

- Les procédures de recherches se doivent de générer un nombre de configurations possibles (énergétiquement optimales) comprenant la configuration cristallographique. Une recherche exhaustive énumèrerait tous les modes d'amarrage entre les deux protéines. Les 6 degrés de liberté (rotation et translation) seraient explorés ainsi que les degrés de liberté internes à chaque molécule. Cela est cependant impossible dû à la taille de l'espace de recherche. Par conséquent, l'espace de recherche est exploré et échantillonné par des méthodes heuristiques moins coûteuses en temps de calcul. La contrepartie est que la solution globale de plus basse énergie peut être manquée (mais celle-ci n'est pas forcément celle choisi biologiquement). De plus, la plupart des programmes d'amarrage s'affranchissent de la flexibilité des protéines mises en jeu (mode de corps rigides). La flexibilité des protéines peut aller du 'simple' mouvement des chaînes latérales aux mouvements de plus grande ampleur de la chaîne peptidique (mouvements allostériques). Elle n'est pas triviale à prédire et pourtant elle joue un rôle non négligeable dans l'interaction par une adaptation mutuelle des protéines (encore nommé 'induced-fit'). Ces mouvements doivent être envisagés dans le cas d'une prédiction qui utilisera it des structures de protéines déterminées de façon séparées.

- La fonction de score doit être capable de distinguer le mode d'amarrage expérimental des autres modes d'amarrage proposés par la procédure de recherche. La fonction de score guide aussi la procédure de recherche vers une solution optimale. Cette fois encore, on utilise des approximations pour accélérer les calculs au détriment de la précision. Les scores les plus précis comme une énergie libre d'association sont hors de notre portée dans le cas présent.

Des sessions équivalentes aux CASP on été créées en 2001 pour évaluer les méthodes de prédiction des interactions entre protéines (sessions CAPRI). Elles se distinguent par leur faible nombre de cibles et de participants comparé aux CASPs (Janin, et al., 2003; Mendez, et al., 2005). La prédiction des interactions entre protéines reste du domaine d'expertise académique et bien loin d'être arrivé à maturité comme l'est la prédiction de structure 3D par homologie ou bien la prédiction des interactions protéine-petite molécule. L'objectif de ces sessions est de faire progresser ce domaine de recherche comme cela a été le cas pour les CASPs.

## Le projet DOCKGROUND

Mon travail de recherche au sein du laboratoire d'Ilya Vakser a été de réaliser une première base de données de complexes protéiques co-cristallisés pour constituer le fondement du projet DOCKGROUND (Douguet *et al.*, *Bioinformatics*, 2006). Ce projet vise à établir un système intégré et dynamique de bases de données dédiées à l'étude et à la prédiction des interactions entre protéines. Il permettra d'améliorer nos connaissances des interactions et de développer des outils de prédiction plus fiables dédiés notamment à la prédiction des interactions entre protéines au sein d'un génome. A ce jour, quatre bases de données, mise à disposition de la communauté scientifique, ont été envisagées :

- une base de données de complexes protéiques co-cristallisés (base '**Bound-Bound**').
- une base de données '**unbound**' de complexes protéiques dont les partenaires ont été cristallisés sous forme monomérique non complexée et sous forme complexée (le complexe protéique existe donc déjà dans la première base de données). Dans le cas où il n'existe pas de structure expérimentale non liée celle-ci sera modélisée par dérivation de la forme complexée à l'aide d'une banque de rotamères (dans un premier temps).
- une base de données de complexes **modèles** dont les partenaires existent dans la première base de données mais qui, ici, vont être représentés par un modèle. Cette base simulera les imperfections des modèles obtenus par une modélisation par homologie par exemple. Les programmes de prédiction doivent être capables de gérer ces approximations et de proposer une configuration/zone d'interaction qui même grossière sera juste.
- Une base de données de complexes '**leurres**', c'est-à-dire des complexes pour lesquels certains programmes de prédictions échouent. Cette base sera ouverte au public pour permettre aux groupes de tester leurs programmes de prédiction. Elle servira à nourrir les jeux tests utilisés pour améliorer les programmes de prédictions.

## La base de données de complexes protéiques 'Bound-Bound' (Douguet, et al., 2006)

### *Contenu*

A ce jour, le nombre de complexes protéiques résolus expérimentalement représente une petite fraction des complexes existants *in vivo*. Il est également plus difficile de déterminer la structure d'un complexe que celle d'une protéine isolée. Cependant, il existe un nombre statistiquement significatif d'exemples qui permet une analyse. La PDB contient actuellement plus de 40000 structures dont ~15000 seraient des complexes biologiques possédant une ASA (surface accessible au solvant) enfouie par chaque protéine d'au moins 250 $Å^2$ (Table 3).

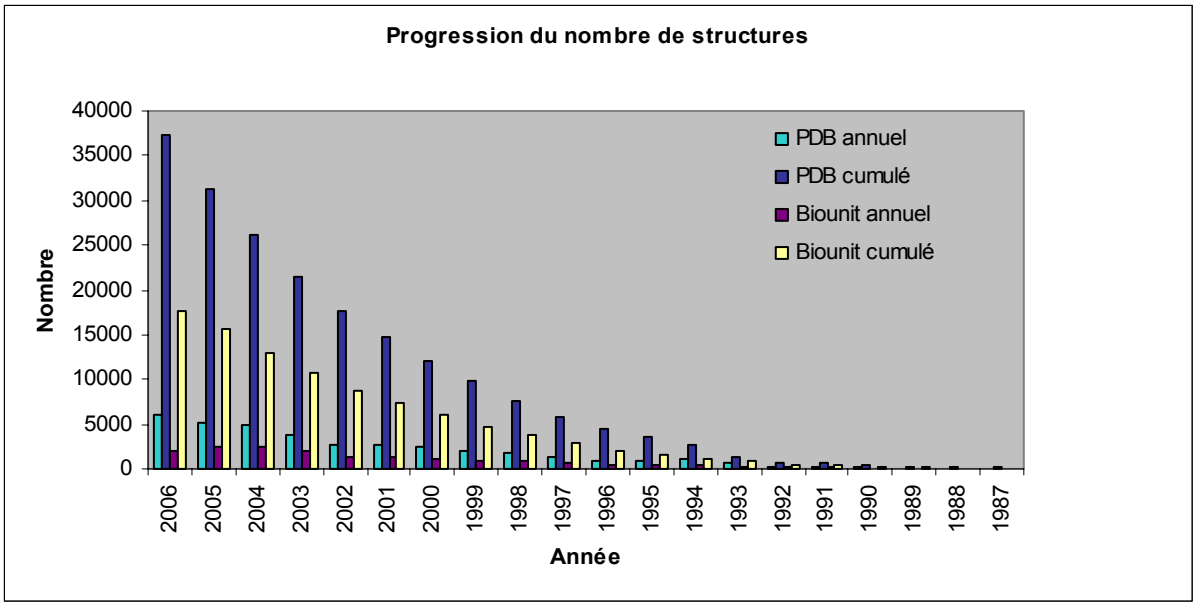



**Table 3**. Progression annuelle du nombre de structures de protéines déterminées expérimentalement et déposées dans la PDB. Le nombre de nouvelles structures accessibles par année est en bleu clair, le nombre cumulé de structures accessibles est en bleu foncé et le nombre de structures 'Biounit' dont au moins 2 chaînes interagissent entre elles est en bordeaux (annuel) et jaune (cumulé) (elles enfouissent une surface accessible au solvant d'au moins 250 $Å^2$ chacune).

Il existe actuellement 3 bases de structures de protéines : la Protein Data Bank (PDB), la Biounit ('Biological Unit File' dérivé de la PDB) et la PQS (Protein Quaternary State de l'EBI également dérivée de la PDB (Henrick and Thornton, 1998)). Une base de données de structures expérimentales est nécessairement fondée sur la PDB d'origine. Dans notre cas, elle se focalise sur le sous-ensemble de structures cristallisées sous forme d'un complexe annoté biologique (ou bien prédit selon la source). Il est ensuite nécessaire que cette base de données soit récente et dynamique étant donné le nombre de complexes protéiques déposés tous les mois (environ 160 depuis 2003). De ce point de vue, les bases dont disposaient la communauté scientifique, en 2004, étaient incomplètes (Bogan and Thorn, 1998; Dasgupta, et

al., 1997; Keskin, et al., 1998; Larsen, et al., 1998; Lijnzaad and Argos, 1997; Lo Conte, et al., 1999; Lu, et al., 2003; Tsai, et al., 1996), laboratoire de Honig (http://honiglab.cpmc.columbia.edu), ou bien avec une mise à jour non automatique et non interrogeables sur l'ensemble des données (Keskin, et al., 2004).

L'utilisation de la PDB originale sous sa forme brute est inappropriée dans la mesure où elle rassemble les fichiers contenant la structure de la ou des protéines présentes dans l'unité asymétrique (ASU) du cristal. L'ASU est la plus petite unité à partir de laquelle on peut reconstruire le cristal dans sa totalité par application d'opérations de symétrie. Les complexes, s'ils existent dans l'ASU, ne sont donc pas nécessairement les complexes biologiques.

Les auteurs de la PQS proposent une base de complexes biologiques créée à partir de l'utilisation d'une procédure automatique qui reconstruit le cristal et qui détermine l'association protéique la plus probable par analyse des contacts entre celles-ci (Henrick and Thornton, 1998).

Enfin, la base Biounit est un sous ensemble de la RCSB PDB, qui contient les fichiers des complexes biologiques tels qu'ils sont définis par les auteurs de la structure (http://pdbbeta.rcsb.org/robohelp_f/data_download/biological_unit/generation_of_biological_unit_coordinate_files.htm).

Nous avons opté pour l'utilisation de cette base de données Biounit comme source sur la configuration biologique probable des complexes co-cristallisés et déterminés expérimentalement. La Biounit et la PQS présentent un certain nombre de différences puisque les résultats de notre comparaison montrent que 19% des entrées pdb communes à la PQS et à la Biounit sont dans une configuration quaternaire différente (37% si l'on effectue le même calcul sur l'ensemble des entrées pdb contenues dans l'une ou l'autre des bases ; résultats non publiés). Vraisemblablement, la PQS et la Biounit contiennent un certain nombre de complexes non biologiques et/ou dans une configuration erronée. Nos résultats ont été confirmés par d'autres études (Jefferson, et al., 2006; Levy, et al., 2006; Xu, et al., 2006).

Les fichiers Biounit ont la particularité d'être conçus sur le modèle des fichiers PDB contenant les structures déterminées par RMN. C'est-à-dire qu'un champ 'MODEL' sépare chaque protéine (encore nommée chaîne) du complexe lorsqu'elle doit être dupliquée pour créer le complexe. Ainsi, elle garde le même nom de chaîne d'origine (la PQS a opté pour une nouvelle dénomination des chaînes dupliquées). Ces particularités compliquent l'identification et le référencement de chaque protéine impliquée dans le complexe car seules les protéines/chaînes contenues dans l'ASU (PDB original) ont un nom unique qui est utilisé par les autres bases de données (information sur la séquence, la classification structurale… comme par exemple dans le système SeqHound qui établit un système de correspondance (Michalickova, et al., 2002)).

Cette collection de fichiers est traitée par un ensemble de programmes qui analysent, annotent et classent les complexes protéiques qui ont été cristallisés. Une structure de la PDB est stockée dans notre base si elle contient un complexe composé de chaînes (segments protéiques) d'au moins 30 acides aminés qui interagissent entre elles.

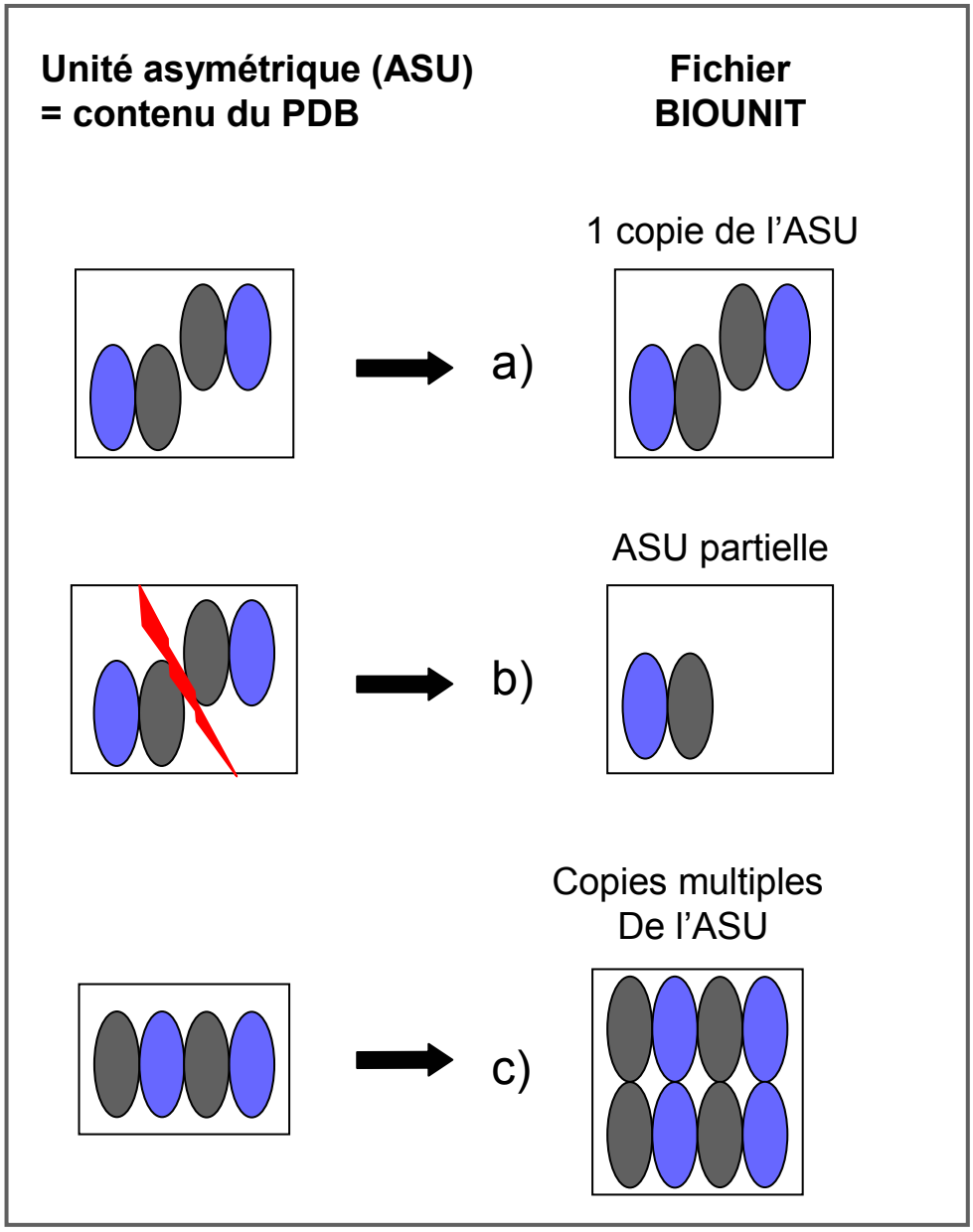



**Figure 9**. L'unité asymétrique (ASU) est le contenu même du fichier PDB d'origine. Afin d'obtenir le complexe biologique (fichier Biounit), il est parfois nécessaire de dupliquer (c) ou de tronquer le contenu de l'ASU (b).

L'interaction est estimée par le calcul de la surface accessible au solvant (ASA) du complexe et celle des chaînes prises séparément (Eisenhaber and Argos, 1993). Lors de la formation d'un complexe, la surface ASA enfouie par chaque protéine impliquée est égale à :

$$ASA = \frac{(ASA_{protéine1} + ASA_{protéine2}) - ASA_{complexe}}{2}$$

Nous avons choisis de représenter les complexes par une, ou des, associations de chaînes prises par paires. La représentation des complexes par paires (une paire équivaut à une interface entre 2 chaînes) permet de stocker d'une façon unique toutes les configurations dans notre base de données. Ainsi, un dimère est représenté par une paire de chaînes (une seule interface) et un trimère peut être représenté par 2 ou 3 paires de chaînes (et donc autant d'interfaces) selon la configuration du complexe. Pour chaque entrée PDB, il est indiqué s'il s'agit d'un homo ou hétéro oligomère et qu'elles sont les interfaces mises en jeu : nom des chaînes impliquées, surface d'interaction, nombre d'acides aminés interagissant.

L'annotation des protéines/chaînes complexées est basée sur plusieurs sources d'information : bases de séquences via le système SRS, l'identification unique du GI (NCBI's GenInfo) par le service SeqHound (Michalickova, et al., 2002), la classification structurale selon SCOP (Hubbard, et al., 1997), la segmentation en domaines de la séquence par ASTRAL (Brenner, et al., 2000) ainsi qu'une évaluation de la qualité de la structure par le score AEROSPACI toujours selon ASTRAL (http://astral.berkeley.edu/). Ainsi, chaque chaîne protéique impliquée dans une interaction est caractérisée par sa séquence en acides aminés, sa fonction via des mots clés et sa classification structurale. Une partie de ce travail a été réalisée par Huei-chi Chen, étudiante à l'Université de Stony Brook (niveau master 1ère année) sous ma direction.

Outre l'intégration des données relatives aux protéines impliquées dans les complexes, il est nécessaire de détecter de façon automatique un certain nombre de particularités non prédictibles ou bien pouvant introduire un biais dans les futures analyses statistiques. Ces complexes particuliers ou 'illégitimes' sont :

- Les faux complexes dans lesquels les deux chaînes qui interagissent sont en fait deux segments contigus d'une même protéine (exemple de pdb2ltn, chaînes A et B, C et D ; figure 10a).
- les complexes emmêlés (exemple de pdb1cma, chaînes A et B ; figure 10b).
- les complexes interagissant uniquement par un segment non structuré (exemple de pdb1fcb, chaînes A et B ; figure 10c).

Il nous a paru nécessaire d'annoter d'autres cas particuliers comme les complexes résultants de la formation d'un pont disulfure ou bien les complexes dont la zone d'interface inclut également un ligand (exemple de pdb1gno), de l'ADN ou de l'ARN (exemple de pdb1gtd).

Enfin, chaque protéine est reliée au code pdb de sa forme monomérique cristallisée si celle-ci existe. Cette détection automatique se base sur le contenu de la Biounit et de l'ASU (fichier pdb d'origine). La protéine est annotée comme étant monomérique si elle est seule dans l'ASU et seule dans le fichier Biounit (donc monomérique selon les auteurs). Cette donnée va faciliter la réalisation de la seconde base de données 'unbound'.

L'ensemble des données est stocké sous la forme d'une base de données relationnelle gérée par PostgreSQL. La procédure de collection, d'analyse, d'annotation et de classification des complexes est automatisée afin de faciliter la mise à jour et de récupérer au plus vite tout nouveau complexe déterminé expérimentalement (figure 11).

La base de données est publique et interrogeable par une interface Web exécutant les requêtes de l'utilisateur (http://dockground.bioinformatics.ku.edu/BOUND/request.php ; figure 12). Le travail sur les requêtes PHP avait été initié par Gaël Lagrange dans le cadre d'un CDD sous ma direction. La page de résultats permet de visualiser la liste des complexes de paires répondant aux critères soumis (par exemple, 1acb E-I ; figure 13a). Pour chaque entrée pdb, un lien mène à la fiche descriptive de la structure et des chaînes la composant (figure 13b). De là, un autre lien mène à la description de l'oligomère ('See the pairwise complexes list' ou ''Multimeric state' ; figure 13c).

La première page de résultats liste les complexes de paires qui répondent aux critères de la requête. Selon ces derniers, un très grand nombre de solutions peut être sélectionné. Il est proposé de réduire la liste des complexes de paires à celle de ses représentants. Les représentants sont essentiels pour éviter une surreprésentation de certaines familles de protéines. Les représentants peuvent être choisis selon leur classification structurale selon SCOP ('Structural Classification Of Proteins') ou bien selon l'identité de séquence choisie par l'utilisateur. Dans ce dernier cas, j'utilise la suite des programmes PISCES (Wang and Dunbrack, 2003) afin de grouper les familles d'une manière dynamique. Au sein de chaque famille, il est possible de choisir le représentant ayant la meilleure résolution cristallographique ou bien le meilleur score AEROSPACI. Ce dernier traduit la qualité globale de la structure après inspection visuelle. Les résultats de la sélection des représentants parviennent à l'utilisateur par l'intermédiaire d'un courriel.

### *Premières analyses des données*

A l'issue de la réalisation de cette base de données, en janvier 2006, 14893 entrées pdb étaient classées comme complexe sur les 34778 que comptait la PDB (actuellement plus de 17000 sur ~ 41000 entrées PDB au total).

- 67 220 complexes de paires (ou interfaces) supérieures à 0 $Å^2$.
- 58% des complexes sont dimériques, c'est-à-dire qu'ils possèdent une seule interface enfouissant au moins 250 $Å^2$ par chaîne.
- 80% des interfaces répertoriées possèdent une ASA enfouie par chaîne d'au moins 250 $Å^2$. La plupart d'entre-elles se situent entre 1000 et 3500 $Å^2$ (seulement 3% au-delà de 3500 $Å^2$).
- 75% des interfaces se font entre deux mêmes protéines (homo).
- 72-74% des chaînes impliquées dans les complexes sont constituées d'un seul domaine selon la définition SCOP (lorsque l'annotation existe).

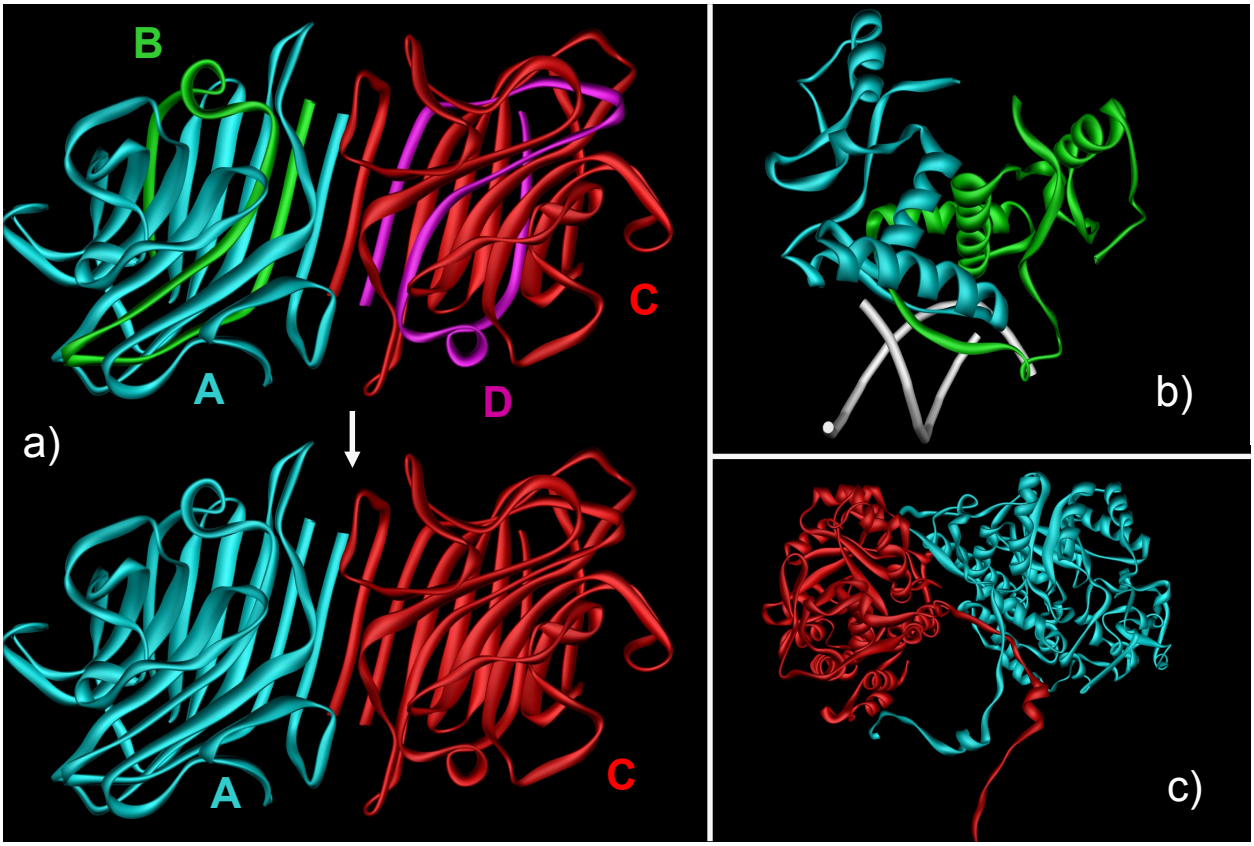



**Figure 10** : Trois exemples de complexes 'illégitimes' automatiquement détectés : a) les chaînes emmêlées et contigues d'une même protéine : 2ltn (AB et CD), b) les chaînes emmêlées : 1cma AB et c) les chaînes interagissant par des segments terminaux non repliées.

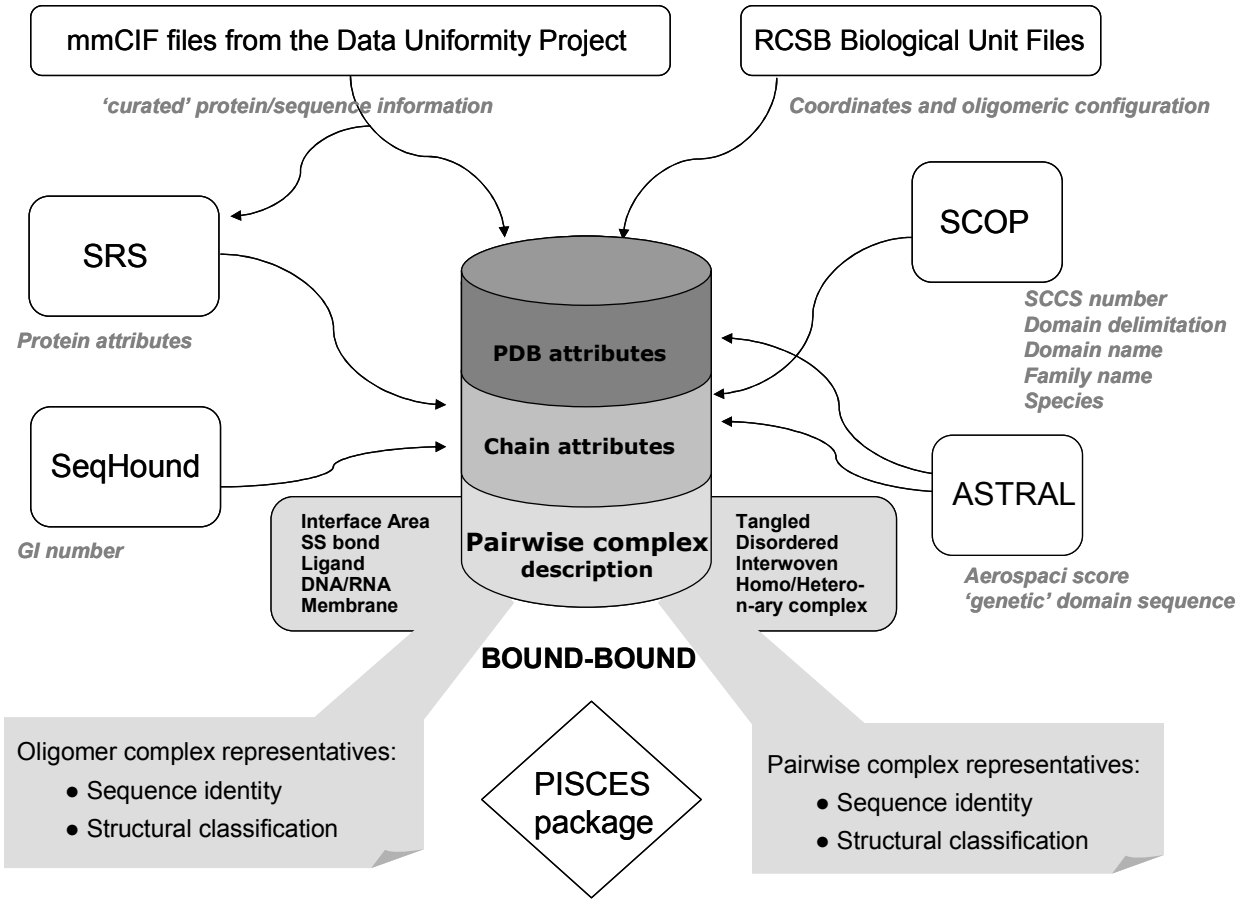



**Figure 11** : Schéma de construction de la base 'Bound-Bound'. Les sources primaires et les programmes externes sont en noir et blanc. La base de données est représentée en gris.

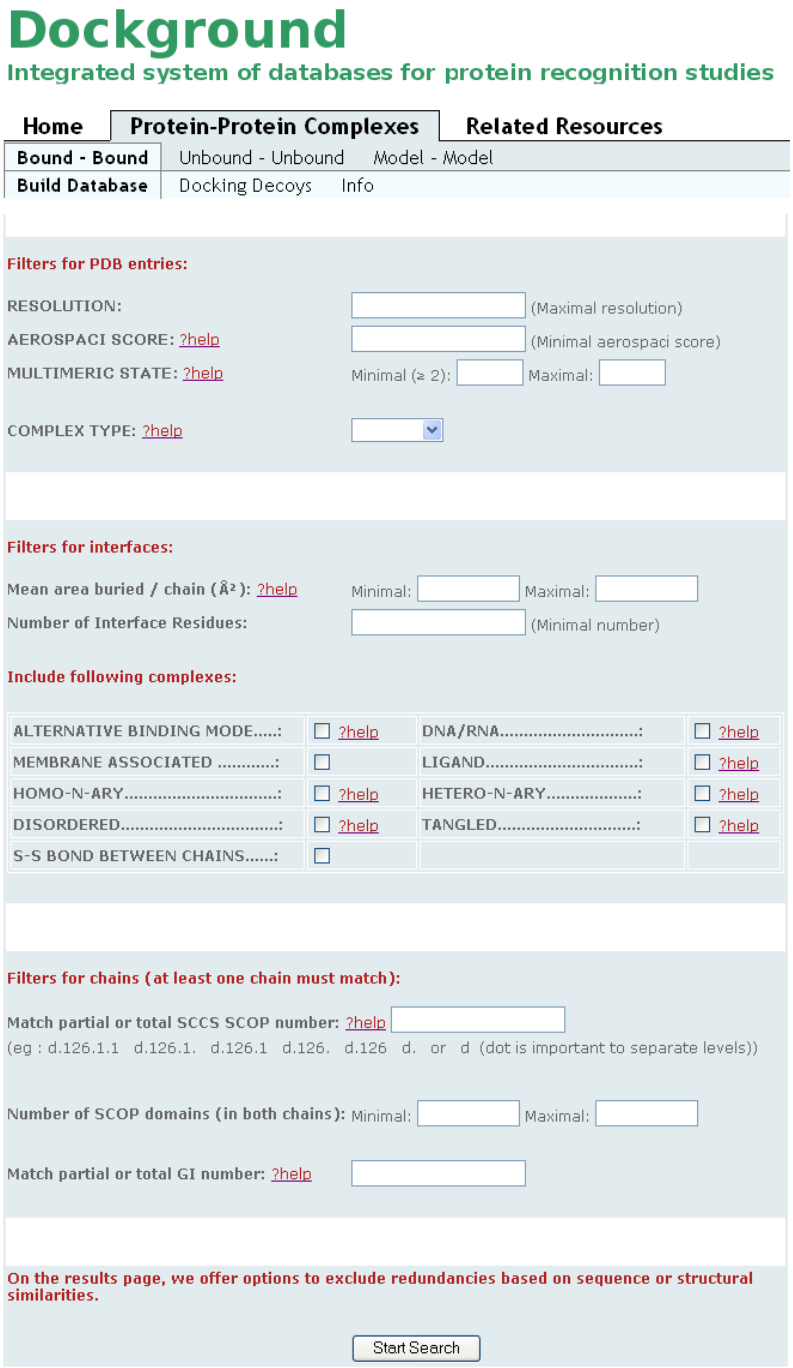



**Figure 12** : Page d'accueil des requêtes sur la base 'Bound-Bound'. a) les requêtes portent sur la structure globale du PDB. L'état multimérique indique le nombre de chaînes formant le complexe avec la contrainte d'une surface d'interaction par chaîne supérieure ou égale à 250 $Å^2$, b) les requêtes portent sur les propriétés et les particularités du complexes, c) la requêtes portent sur les chaînes constituant le complexe.

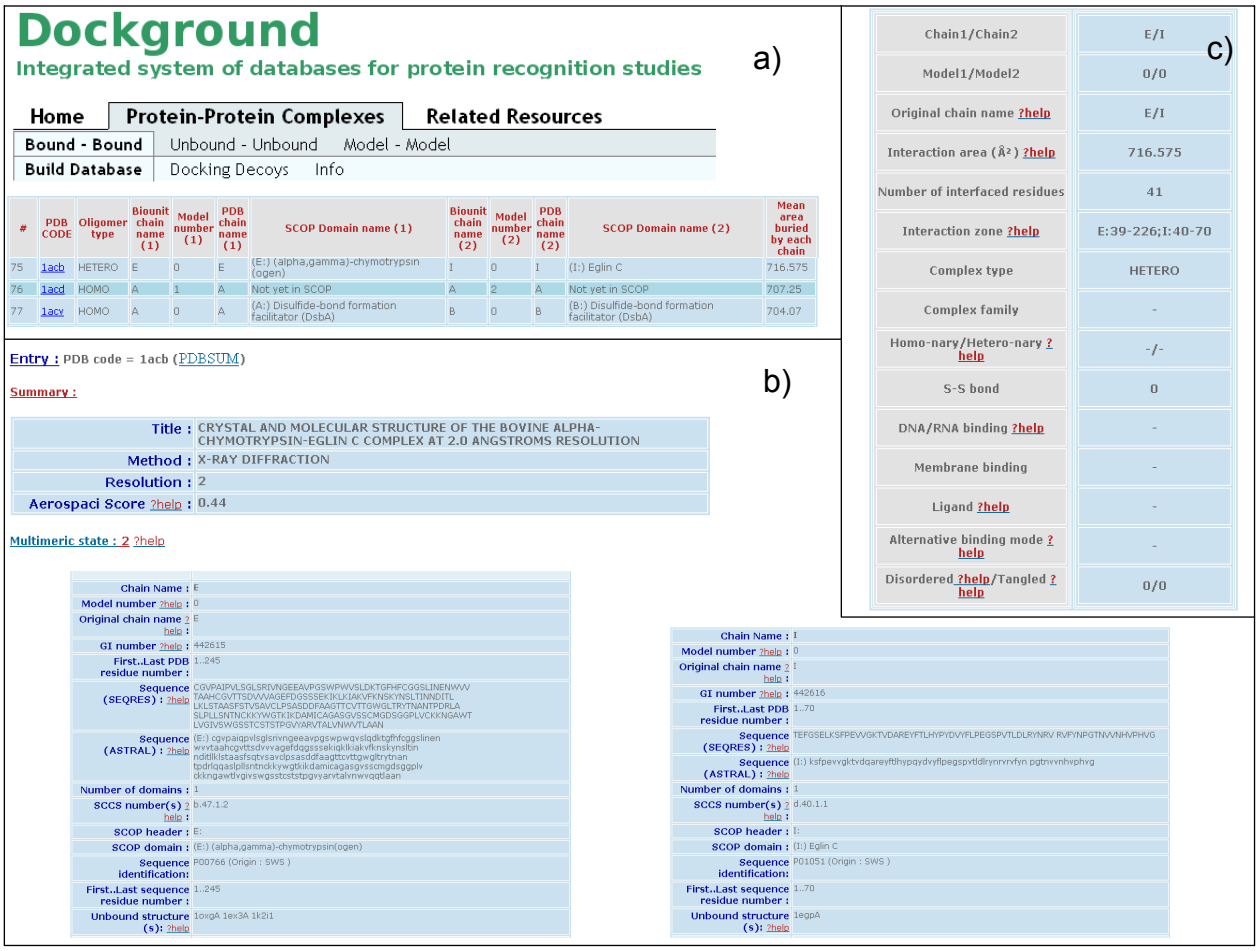



**Figure 13** : Exemple de page de résultats. a) en haut à gauche, la liste des complexes de paires répondant aux critères soumis ; b) en bas à gauche, le détail du contenu du complexe PDB 1acb, et c) le descriptif de l'interface du dimère formé par les chaînes E et I.

***Diversité des complexes***

Dans l'exemple suivant, nous avons tout d'abord sélectionné 4479 complexes de paires sur la base de : entrées non obsolètes, dimériques d'interface $\geq 500$ Å$^2$ et ne possédant aucune des particularités citées précédemment. Deux modes de sélection permettent soit de sélectionner les représentants des complexes de paires de la liste créée soit de sélectionner les représentants des oligomères présents dans la liste. Dans ce dernier cas, les complexes de paires associés aux entrées pdb de la liste sont ajoutés (92 paires supplémentaires dans cette analyse). Ceux-ci ne répondent pas aux critères de sélection mais ils sont importants si l'on souhaite ne sélectionner que les vrais dimères (entrées PDB ne possédant que deux chaînes).

- Dans le mode par paires, nous obtenons 1476 représentants et autant d'entrées PDB ne partageant pas plus de 30% d'identité de séquence. En terme de famille structurale, 960 familles sont représentées (numéro SCOP différents).

- Dans le mode oligomère, il y a 1575 représentants ne partageant pas plus de 30% d'identité de séquence dont 1460 sont les représentants des vrais dimères. Les homodimères représentent 82% de ces représentants. Il n'existe seulement que 261 cas d'hétérodimères.

Ce jeu de données pré compilé est téléchargeable sur le site Web de DOCKGROUND sous la forme d'un fichier texte lisible par Excel ('Easy mode'). Un second jeu de donnée a été créé manuellement et mis à disposition du public. Dans ce jeu, les cas où deux sous unités d'une même protéine interagissent obligatoirement ont été éliminés. Ce jeu contient aussi des cas de complexes homologues qui partagent plus de 30% d'identité de séquence mais qui présentent un mode d'amarrage diffèrent. Il contient environ 500 complexes classés enzyme-inhibiteur, antigène-anticorps, cytokine/récepteur ou 'autres'.

Notre base de données est la seule, à ma connaissance, à permettre une recherche dynamique de complexes par des critères aussi variés (caractéristiques des interfaces, mais aussi par classe/repliement/superfamille ou famille SCOP) et à générer une liste de représentants si cela est nécessaire.

Des améliorations sont à apporter dans le futur comme par exemple des précisions sur la localisation des molécules d'eau. Des travaux récents, non encore publiés, leur accordent une réelle importance (F. Cazals, INRIA, Sophia-Antipolis). Ce problème n'est pas trivial avec l'utilisation des fichiers Biounit dans lesquels une interface (et donc la position exacte des molécules d'eau) n'est pas toujours observée dans l'ASU.

***Prédictions par homologie ?***

Dans l'hypothèse d'une transférabilité du mode structural d'interaction entre protéines auquel on applique le principe de la prédiction par homologie à un taux d'identité de séquence de 30%, on réalise combien les données expérimentales, 261 cas d'hétérodimères seulement, sont loin de couvrir l'espace des associations possibles entre protéines. Par comparaison :

- Modélisation par homologie d'une protéine :
  ~40 000 structures expérimentales PDB
  ~6 000 familles de structures supports (source PISCES à 30%)
  ~21 000 modèles chez l'humain (source MODBASE ~60% du génome)
  (Pieper, et al., 2006)

- Modélisation par homologie des interactions (cas du dimère, le plus représenté) :
  ~40 000 structures expérimentales PDB
  ~17 000 complexes
  ~1 460 familles de dimères à 30%
  ~260 familles d'hetéro-diméres à 30% pouvant servir de support
  ~ ? modèles

L'utilisation de la transférabilité structurale est clairement limitée par le nombre d'exemples disponibles surtout dans le domaine des hétéro-oligomères. Il faudra de nombreuses années avant d'obtenir un répertoire structural complet des complexes existants dans une cellule. L'hypothèse est également malmenée par les expériences de mutagenèse qui montrent parfois qu'une seule mutation peut entraîner un réarrangement majeur tant dans la structure tertiaire que quaternaire (exemple de LicT PDB1tlv versus PDB1h99). Il est important d'être prudent lorsque l'on veut transférer la configuration d'un complexe à un autre membre de la famille. Par exemple, les domaines LIM diffèrent par leur mode de liaison (LIM1/LIM4) et 2 mutations sur LIM4 inhibent l'interaction avec SH3 (Velyvis, et al., 2003). Pourtant, une étude de 2003 va dans ce sens (Aloy, et al., 2003). Et malgré le peu d'exemples de complexes déterminés expérimentalement certains groupes s'aventurent déjà à prédire les interactions entre protéines au sein d'un génome (Aloy, et al., 2004; Davis, et al., 2006; Korkin, et al., 2006; Lu, et al., 2003; Russell, et al., 2004). Chez la levure, ils obtinrent un recouvrement de 3% avec les données expérimentales de la base BIND (Alfarano, et al., 2005).

D'autres alternatives bioinformatiques, de moindre résolution, prédisent les sites potentiels d'interaction : alignement de séquences, arbres phylogénétiques (Lichtarge, et al., 1996) ou identification des 'hot spots' (Ma, et al., 2003). Par ailleurs, les méthodes de prédictions développées jusqu'à présent s'apparentent davantage aux méthodes d'amarrage ('docking') de petites molécules. Elles sont basées sur des fonctions de score empiriques (potentiels de Lennard-Jones par exemple), sur des potentiels statistiques ainsi que toute autres contraintes permettant de réduire l'espace de recherche (information biologique issue de la littérature par exemple).

Les méthodes hybrides sont cependant les plus prometteuses. Il s'agit de combiner l'information structurale des protéines isolées avec les informations structurales sur l'enveloppe du complexe qu'elles forment (méthodes expérimentales de basse résolution comme la cryo-microscopie électronique (~5nm) ou le SAX ; (Zhou, et al., 2001)). A défaut de données cristallographiques sur les protéines isolées, celles-ci peuvent être modélisées par homologie. L'objectif est de déterminer la configuration du complexe la plus probable. Celle-ci peut être validée par mutagenèse dirigée par exemple.

## Conclusions

Certains complexes resteront difficiles à déterminer tel que les éphémères (encore nommés 'transient') majoritairement composés d'hétéro-complexes, *a contrario* des obligatoires ('obligate') plus stables, majoritairement homo-oligomèriques et donc plus facilement observables. De plus, l'exemple du complexe LIM4/SH3 montre bien qu'une interaction peut être à la fois d'une affinité faible (mais observable par RMN) mais très spécifique ; il suffit de deux mutations de type Arg→Ala pour inhiber la formation du complexe (fait confirmé par l'absence d'interaction avec un homologue ; (Velyvis, et al., 2003)). Des résultats similaires ont été obtenus sur l'interaction TEM1/BLIP (Reichmann, et al., 2005). Ces derniers innovent dans la description des interfaces en les caractérisant par des clusters de résidus dont le nombre et la densité des connectivités (interactions : intra (effet coopératif) et inter (effet additif)) et, non l'étendue de leur surface, peut expliquer l'affinité d'un complexe. Au-delà d'une complémentarité de surface, une interface est donc mieux caractérisée en terme d'organisation interne. En terme de spécificité, la mutation d'un seul résidu peut davantage déstabiliser toute l'organisation que de muter 5 au sein du même cluster. Une étude plus représentative serait probablement très instructive. D'autres équipes en collaboration avec des experts en géométrie sont en cours pour mieux définir ce qu'est une interface et comment la caractériser (Cazals, et al., 2006). Leur premiers résultats bousculent déjà quelques idées reçues (serveur http://bombyx.inria.fr/Intervor/intervor.html ; publication en cours).

En complément à ce chapitre, voici une liste de serveurs traitant des interfaces protéine-protéine selon d'autres perspectives :

Bases de données s'organisant autour des interactions domaine/domaine observées dans la PDB :

- La base de données PIBASE utilise la PQS comme source de ces complexes et elle est interrogeable par le code PDB ou le numéro SCOP ou CATH (http://alto.compbio.ucsf.edu/pibase/queries.html ; (Davis and Sali, 2005)).
- 3did est une base de données organisée autour des domaines PFAM et interrogeable par nom de domaine (http://3did.embl.de ; (Stein, et al., 2005)).
- iPfam est une base de données organisée autour des domaines PFAM (http://www.sanger.ac.uk/Software/Pfam/iPfam/ ; (Finn, et al., 2005)).
- SCOWLP est une base de données organisée autour des domaines SCOP (http://www.scowlp.org/ ; (Teyra, et al., 2006)).
- PSIMAP, PSIBASE et InterPare sont des bases de données organisées autour des domaines (PFAM, SCOP, FSSP, CATH) (http://interpare.kobic.re.kr ; (Gong, et al., 2005)).
- SCOPPI permet une recherche par numéro SCOP et en affiche les alignements multiples (http://www.scoppi.org/ ; (Winter, et al., 2006)).

Bases orientées 'structure' :

- Une base de classification structurale avec une description de la topologie des complexes. Elle est basée sur la Biounit (www.3Dcomplex.org ; (Levy, et al., 2006)).
- La base de données ProtBuD a pour objectif d'identifier le complexe support pour une modélisation par homologie. Elle est basée sur la Biounit (http://dunbrack.fccc.edu/ProtBuD/ProtBuD.php ; (Xu, et al., 2006)).

## IV - Identification de molécules bio-actives par criblage virtuel et *de novo* 'drug design'

Depuis Novembre 2002, je suis rattachée au Centre de Biochimie Structurale de Montpellier en tant que chercheur INSERM. L'objectif de mes travaux est d'identifier de nouveaux composés bio-actifs afin de comprendre le fonctionnement de leur cible dans un contexte biologique et d'évaluer leur 'drogue-abilité', plus particulièrement dans le domaine de l'anti-infectieux (*Mycobacterium tuberculosis*) et de l'oncologie.

Les stratégies actuelles de recherche et de développement de médicaments s'appliquent à réduire le temps affecté à la découverte d'une nouvelle drogue. Ces procédures misent sur le haut débit par l'automatisation des tests *in vitro*, le développement de la chimie combinatoire et l'utilisation de la bioinformatique et de la modélisation moléculaire (Figure 14 ; (Walters, et al., 1998)). La modélisation moléculaire s'est adaptée au haut débit en s'orientant vers le criblage *in silico* (virtuel) de très nombreuses molécules (Shoichet, 2004).

Dans le passé, la chimie médicinale se basait sur les résultats des tests *in vitro* et/ou *in vivo*, sur la structure des substrats et sur une modification par incrément des ligands connus pour identifier de nouveaux 'candidats-médicament'. Par la suite, le succès de la cristallographie par rayons X a permis l'élaboration de médicaments en se basant sur la structure 3D de la protéine cible (exemples du Viracept, un inhibiteur de protéase du HIV (Wlodawer and Vondrasek, 1998) ou du Relenza, un anti-influenza (von Itzstein, et al., 1993)). Plus récemment, l'introduction du criblage virtuel a offert une nouvelle voie d'identification de ligands. Le criblage virtuel basé sur la structure 3D de la protéine consiste à amarrer et à prédire l'affinité d'un grand nombre de molécules (collectées en chimiothèques) pour le site actif ciblé. Les molécules les plus prometteuses sont sélectionnées, achetées ou synthétisées puis testées expérimentalement. Le criblage virtuel permet d'accéder à un espace des molécules beaucoup plus grand que celui qui serait exploré par l'utilisation de techniques expérimentales (en raison du coût, du temps et de l'infrastructure nécessaire à ces dernières). Il permet une présélection qui augmente la probabilité d'identifier un ligand (par exemple : 34.8% contre 0.021% de molécules actives pour le criblage virtuel et HTS respectivement (Doman, et al., 2002) ; Figure 15). De plus, les informations sur le mode d'amarrage d'un ligand permettent de comprendre structuralement son mode d'action et d'envisager son amélioration par dérivation. Ce type de criblage par assemblage ('docking') nécessite, bien entendu, la connaissance des données structurales, expérimentales ou modélisées, de la protéine cible. Le criblage virtuel a été appliqué avec succès sur un certain nombre de protéines dont la structure 3D en complexe avec le ligand a été parallèlement déterminée (anhydrase carbonique humaine (Gruneberg, et al., 2002) ou encore l'AmpC β-lactamase (Gruneberg, et al., 2002; Powers, et al., 2002). D'autres succès au cours des années 1999-2001 sont répertoriés ($Ca^{2+}$ antagonist/T-channel blocker, FKBP ligand, Thrombin inhibitor, K+ channel (kv 1.5) blocker, Farnesyltransferase inhibitor, Glyceraldhehyde-3-phosphate dehydrogenase (GAPDH) inhibitor, Aldose reductase (AR) inhibitors, HIV-1 RNA transactivation response element (TAR) inhibitor (Schneider and Bohm, 2002) ; DNA gyrase inhibitor (Boehm, et al., 2000)).

Plus récemment (2005), la société Astex a obtenu l'autorisation d'entreprendre les tests cliniques sur le composé AT7519 comme anti-cancéreux, 14 mois seulement après sa synthèse. Actuellement, cette molécule est en développement clinique phase I/IIa. Cette molécule a été identifiée par une stratégie qui combine la chémoinformatique et les criblages structuraux expérimentaux basés sur les fragments.

Les méthodes que j'exploite appartiennent au domaine de la chémoinformatique et de la modélisation moléculaire. Cela inclu la modélisation par homologie de la structure des protéines ciblées telle qu'elle a été présentée au chapitre 1, le criblage virtuel de chimiothèques et le *de novo* 'drug design'.

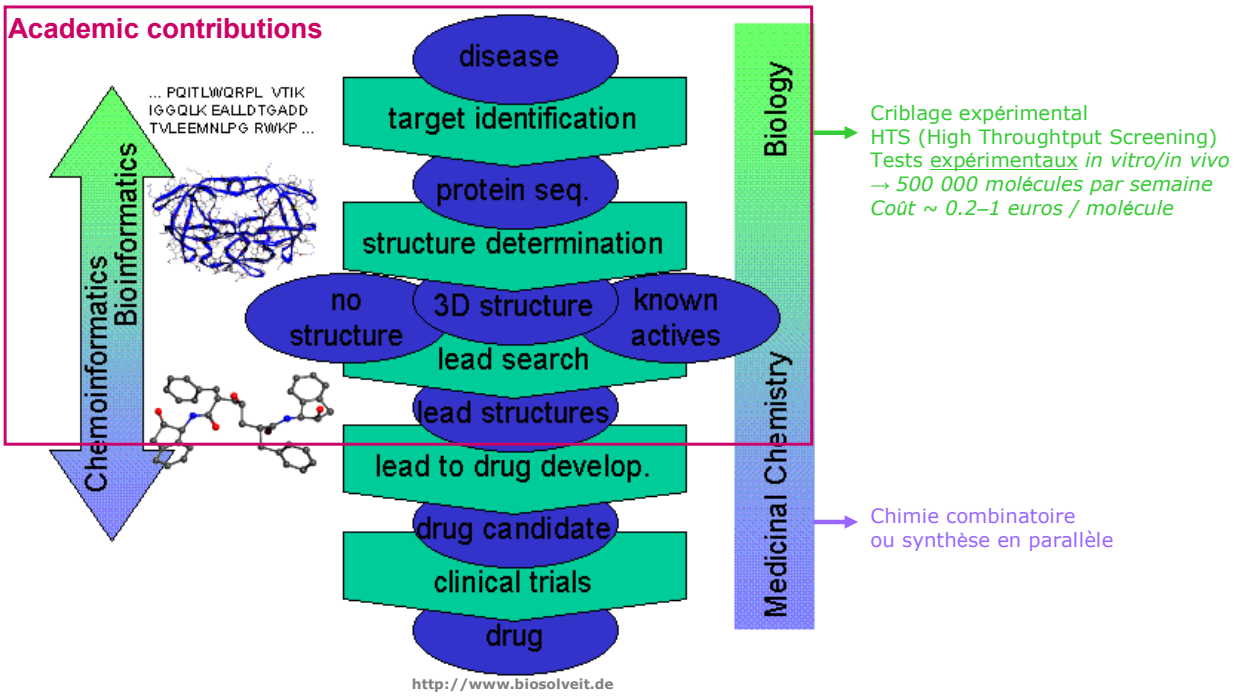



**Figure 14** : Schématisation de la procédure du développement d'une drogue par l'industrie pharmaceutique sur 10-14 années. Les activités de recherche dans un contexte académique sont centrées sur les premières étapes. La bioinformatique et la modélisation moléculaire peuvent intervenir dès l'identification de la protéine d'intérêt.

a)

**Objectif: optimiser le rapport** $\dfrac{\text{Nombre de molécules actives}}{\text{Nombre de molécules testées}}$

b)

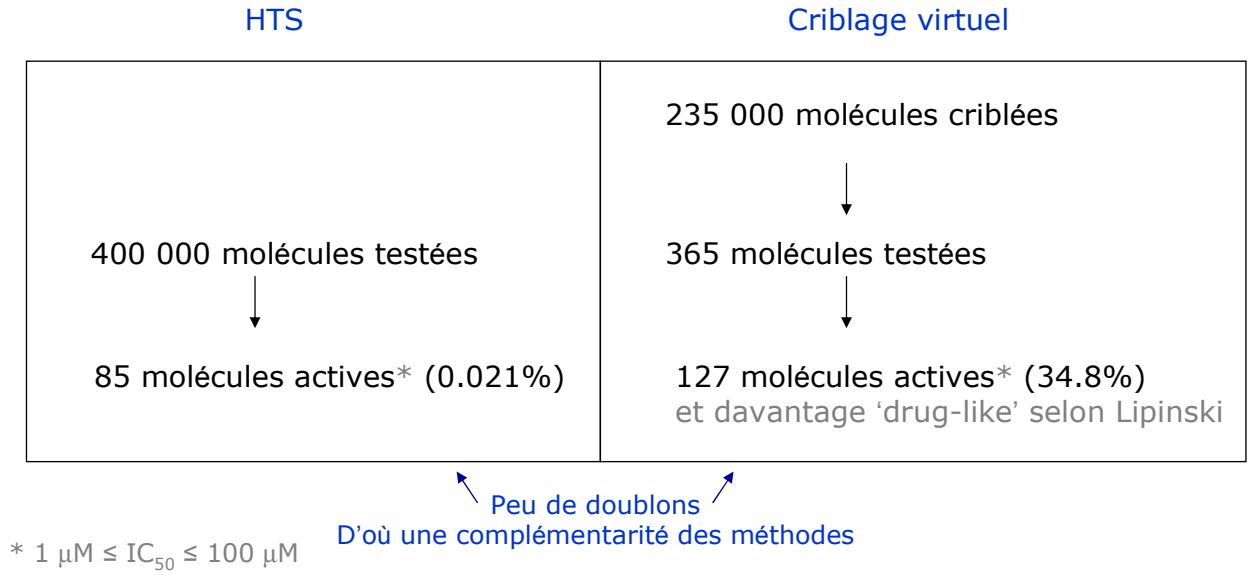



**Figure 15** : a) Le criblage virtuel a pour objectif d'optimiser le rapport entre le nombre de molécules effectivement actives sur le nombre de molécules testées. b) Comparaison des criblages virtuels et expérimentaux HTS sur la Protein Tyrosine Phosphatase B (PTP1B) étudiée pour son implication dans le diabète de type II. Les résultats ont été publiés par la société Pharmacia acquise depuis par Pfizer. Les résultats montrent que le criblage virtuel permet un enrichissement important en molécules actives de la chimiothèque testée. Cependant, les résultats indiquent aussi que les molécules identifiées par l'une ou l'autre des méthodes sont de classes structurales différentes d'où la complémentarité des méthodes.

La conception et la mise au point de certaines des méthodes fait partie intégrante de mon travail de recherche. Elles sont complémentaires des programmes académiques déjà existants. Les outils développés me permettent d'exploiter des nouvelles stratégies d'identification de ligands, de travailler, à la carte, sur une cible afin d'optimiser la recherche (intégration de données diverses) et enfin de m'affranchir de licences coûteuses pour des programmes commerciaux assimilables parfois à des 'boîtes noires' et créés pour traiter le cas général.

Plus particulièrement, je développe depuis ma thèse un programme de *de novo* 'drug design' LEA3D (pour 'Ligand by Evolutionary Algorithm' 3D) (Douguet, et al., 2005) dérivé de la première version de LEA conçue durant ma thèse (Douguet, et al., 2000). Le programme génère des petites molécules à partir de la combinaison de fragments moléculaires issus de drogues et de molécules 'bio' (substrats ou produits de réactions enzymatiques). La stratégie d'optimisation de la combinaison par un algorithme génétique s'inspire de la théorie de l'évolution développée par C. Darwin. On fait évoluer un ensemble de molécules sur plusieurs générations par des opérateurs de croisement et de mutation. La procédure itère jusqu'à ce qu'apparaissent des molécules parfaitement adaptées aux contraintes demandées (Figure 16). Une contrainte est, par exemple, l'affinité prédite d'une molécule pour le site actif d'une protéine. Ce score associé à chaque molécule candidate est le même que celui qui est calculé lors d'un criblage virtuel. Le *de novo* 'drug design' et le criblage virtuel sont deux modules que l'on peut combiner de façon très intéressante. Le premier pour imaginer les molécules 'idéales' desquelles on extrait le pharmacophore 'idéal' et le second pour tester les molécules commerciales disponibles qui possèdent ce pharmacophore. Un pharmacophore est un arrangement de fonctions ou groupes chimiques dans l'espace responsable de l'affinité (prédite ou réelle) d'un ligand pour sa cible protéique. LEA3D permet de gérer les criblages virtuels en utilisant simplement les deux premiers modules (lecture de la chimiothèque initiale et évaluation du score des molécules présentes dans la chimiothèque).

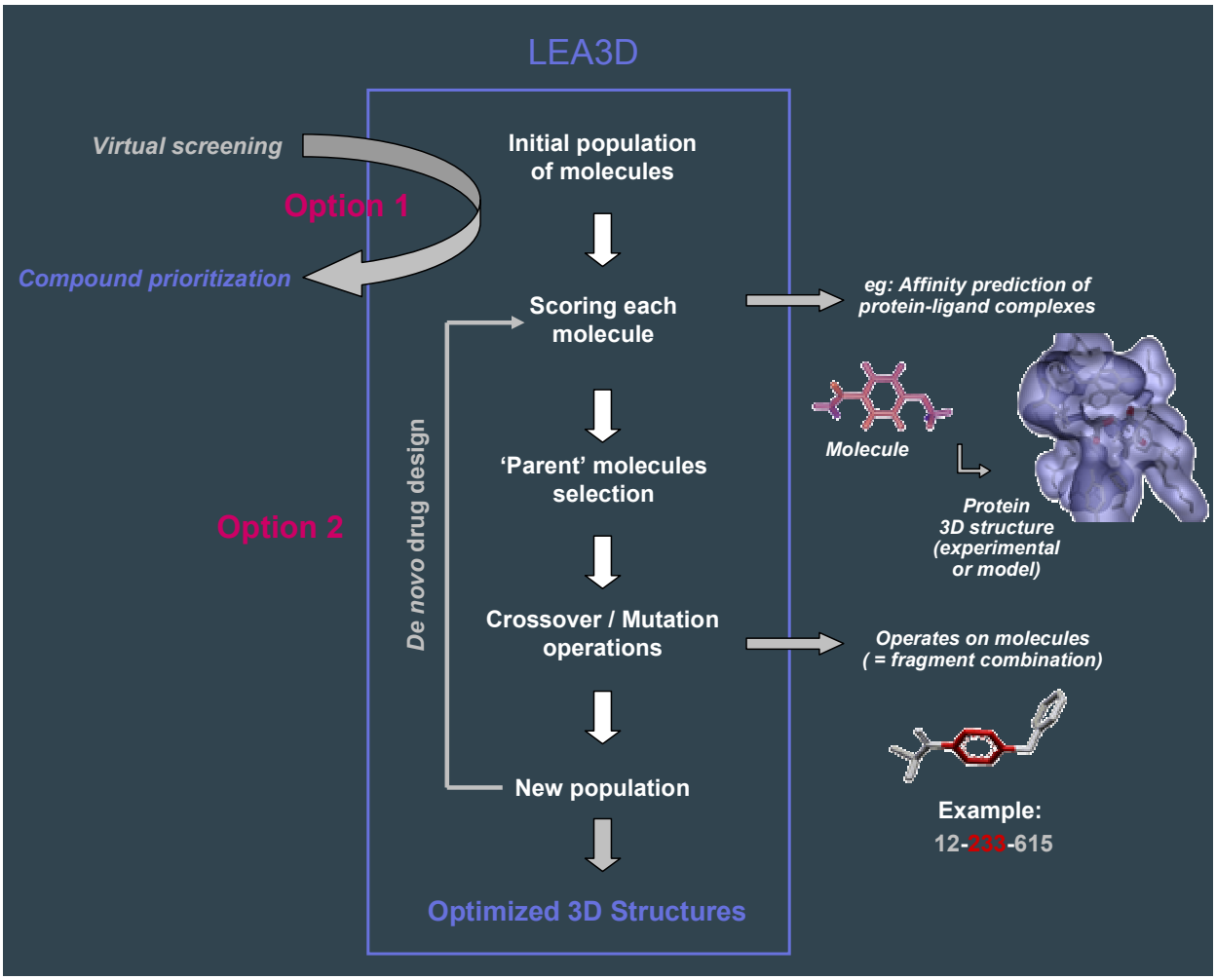



**Figure 16** : Présentation des 2 modules principaux de LEA3D. Option 1 : le criblage virtuel d'une chimiothèque de départ afin de donner la priorité aux composés les plus prometteurs. Option 2 : le *de novo* 'drug design' crée de nouvelles molécules de structure 3D optimisée à partir de la combinaison de fragments. Ces fragments proviennent de drogues (source : base 'Comprehensive Medicinal Chemistry') et de molécules substrats ou produits de réactions enzymatiques (source : GenomeNetJapan). La fonction de score est modulable : évaluation des propriétés physico-chimiques, prédiction de l'affinité pour une protéine…

**Identification d'inhibiteurs de la thymidine monophosphate kinase (TMPK) de *Mycobacterium tuberculosis* par criblage virtuel et *de novo* 'drug design'**

Ce cas d'étude est le plus ancien et le plus abouti depuis mon recrutement à l'INSERM. Il fait intervenir une biologiste, une chimiste, un cristallographe. Mon travail d'identification de molécules bio-actives par chémoinformatique intervient à différents stades d'un projet. Cela peut être l'identification des premiers ligands d'une protéine (même à partir d'un modèle), la modification de ligands connus, l'optimisation de certaines propriétés comme la spécificité… Chaque cas d'étude doit être traité spécifiquement. De plus, dans le cadre du criblage de molécules sur la structure 3D d'une protéine, le post-traitement des résultats est une étape très importante qui doit être réalisée avec soin en employant toutes les connaissances disponibles sur la dite protéine pour affiner, automatiquement et manuellement, la sélection.

Comme cela a été indiqué plus haut, le criblage virtuel consiste, par exemple, à amarrer et à prédire l'affinité de petites molécules pour le site actif d'une protéine dont on connaît, de préférence, la structure 3D. Cette méthodologie (via le programme commercial FlexX), et le *de novo* 'drug design' par LEA3D, a été appliquée avec succès à la Thymidine MonoPhosphate Kinase de *Mycobacterium tuberculosis* (TMPKmt) dans le cadre d'une collaboration avec une chimiste et une biologiste de l'Institut Pasteur (Sylvie Pochet et Hélène Munier-Lehmann, respectivement).

La TMPKmt appartient à la famille des nucléosides monophosphates kinases qui catalyse le transfert réversible d'un groupe phosphate de l'ATP à un nucléoside monophosphate. Cette protéine est essentielle à la croissance de la mycobactérie. C'est la raison pour laquelle elle est étudiée depuis quelques années (Fioravanti, et al., 2003; Li de la Sierra, et al., 2001; Pochet, et al., 2002).

De nouvelles familles d'inhibiteurs (Douguet, et al., 2005; Vanheusden, et al., 2004) ont été identifiées dont un inhibiteur synthétique trois fois plus afin que le substrat naturel ($Ki_{Mol969\_UMR7081}$=1.5 µM versus $Ki_{dTMP}$=4.5 µM ou encore $Ki_{dT}$=27 µM). Les résultats du criblage virtuel n'ont pas été publiés (Figure 18). Avec un taux de réussite de plus de 40%, ces résultats sont bons et valident notre modèle et notre approche. Ils nous ont encouragés à poursuivre un projet de *de novo* 'drug design' sur de nouveaux analogues de la thymidine. Plus particulièrement, l'accent a été mis sur une série d'inhibiteurs dont le précurseur avait été proposé par le *de novo* 'drug design' (Douguet, et al., 2005). Dans ce travail, on montre que le ribose de la thymidine peut être avantageusement remplacé par un groupe benzyl (Figure 17). Cette démonstration s'est tout d'abord appuyée sur l'achat d'analogues commerciaux afin de limiter les étapes de synthèse coûteuses en temps. Le benzyl simple (sans substituants) possède déjà à lui seul une activité de 75 µM et lorsqu'il est substitué par un para-4-acide butanoïque, le Ki est de 13 µM.

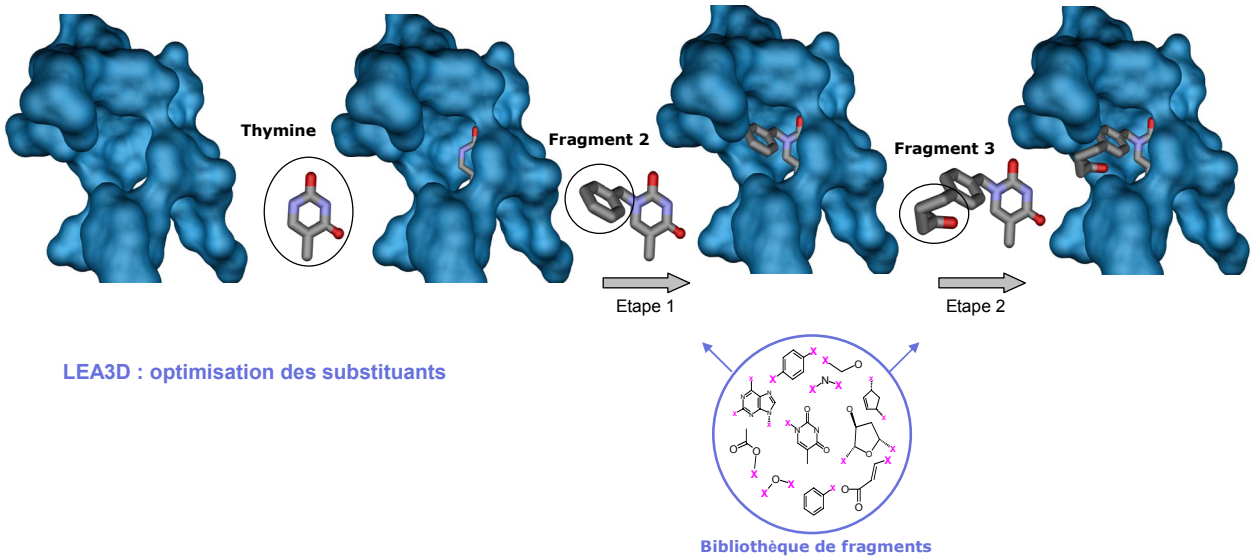
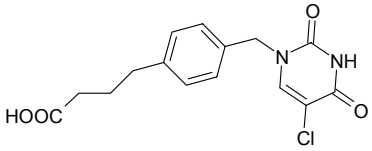



**Figure 17** : Optimisation des substituants de la nucléobase thymine par LEA3D en deux étapes.

La série s'est diversifiée par la synthèse de molécules (~ 40) dont on a fait varier les substituants en position 5 de la nucléobase et en position para du benzyl (publication soumise à *J. Med. Chem.*). Aujourd'hui, le meilleur inhibiteur de cette famille (le 4-[4-(5-Cl-dU)phenyl]-butyric acid) possède un Ki de 6.5 µM, plus faible que le Ki du dTMP (désoxythymidine monophosphate) mais il a l'avantage d'une spécificité vis-à-vis de la TMPK humaine plus importante (Ki=790 µM ; index de sélectivité de 120). Son activité (MIC50) sur *M. bovis* (BCG) est de 45 µg/mL. Cette dernière reste modérée si on la compare à celle de l'izoniazide utilisée comme anti-tuberculeux standard (INH ; MIC50=0.05 µg/mL).




acide 4-[4-(5-Cl-dU)phenyl] butyrique
Ki = 6.5 µM

L'optimisation des substituants du benzyl est pourtant loin d'être achevée et elle nécessiterait bien plus de moyens pour la chimie (amélioration de la complexité (et de l'affinité) et de la spécificité vis-à-vis de l'homologue humaine). La figure 19 indique le mode de fixation prédit pour le meilleur analogue synthétisé. Bien entendu, seule la diffraction des rayons X sur le co-cristal avec la protéine apporterait la preuve de ce mode de fixation. Mais jusqu'à présent, il nous a été impossible de reproduire les conditions de cristallisation publiées. On se base donc sur les mesures d'inhibition compétitive et les relations structure-activité pour valider nos hypothèses.

Les molécules citées sont en cours de protection par un brevet (demande PCT Europe, Canada, Inde et Etats-Unis) pour leur activité antimicrobienne d'une façon générale laissant, aussi, la possibilité de les exploiter pour leur activité sur d'autres cibles protéiques fixant des dérivés de la nucléobase thymine.

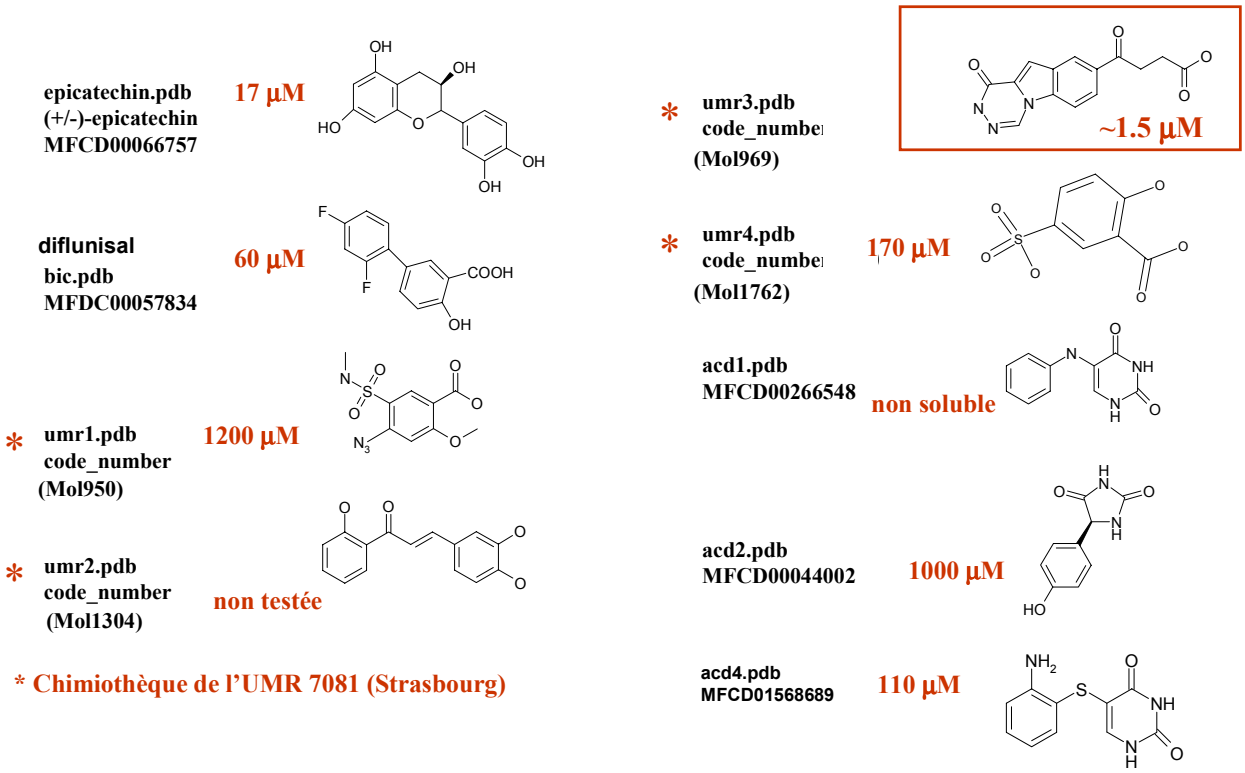



**Figure 18** : Sélection de 9 molécules par notre approche du criblage virtuel combiné à la sélection de pharmacophores par *de novo* 'drug design'. La chimiothèque de départ contenait environ 600 000 molécules, la recherche par pharmacophore l'a réduite à ~3000 composés et le criblage final couplé à la sélection manuelle en a retenu 9. 7 molécules ont été testées (l'une n'était plus disponible et l'autre était insoluble). 4 molécules ont un Ki ≤ 110 µM soit un ratio molécule actives sur molécules testées de 57% (ou 42% si on prend les 3 premières de Ki ≤ 100 µM comme dans la figure 14).

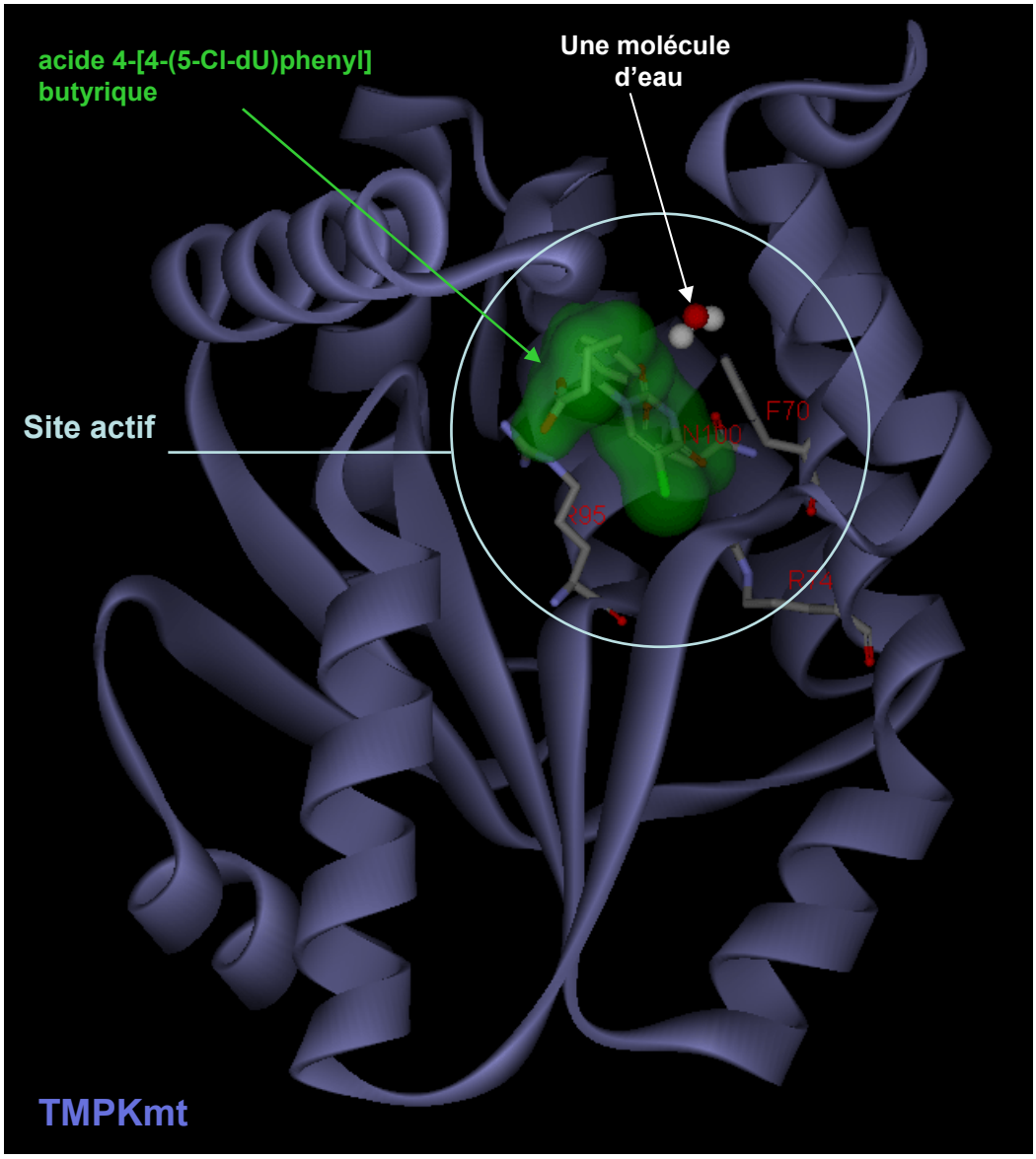



**Figure 19** : Représentation du mode de fixation prédit par le programme FlexX entre la thymidine monophosphate kinase de *Mycobacterium tuberculosis* et l'inhibiteur synthétisé le plus affin. Le ligand est représenté avec sa surface accessible au solvant en vert. Quatre résidus importants pour la fixation du substrat sont indiqués en rouge : R74, F70, N100 et la molécule d'eau fixent la base thymine commune aux substrats (dTMP et dT) et aux analogues synthétisés. Le quatrième résidu R95 est conservé au sein des TMPKs homologues (humain, *Escherichia coli*, *Bacillus subtilisis*, *Yersinia pestis* et *Haemophilus influenzae*). Il est important pour la fixation des substrats et prédit important pour la fixation de notre analogue.

## Identification de molécules bio-actives basée sur le criblage de fragments

Le nombre de drogues mises sur le marché est très faible comparé à l'espace des molécules théoriquement synthétisables. La majorité, 50 à 80 %, de cet espace est inexploitable (Baurin, et al., 2004). Cependant, on peut considérer que si l'objectif est en premier lieu d'étudier la fonction biologique de la protéine ciblée, alors les ligands identifiés ne doivent pas nécessairement posséder toutes les caractéristiques ADME-T associées à une drogue. L'avantage est d'étendre l'espace des molécules exploitables. Mais il faut garder à l'esprit qu'une molécule bio-active non 'drug-able' pourra poser des problèmes lors du passage des tests réductionnistes (exemple : un simple test d'affinité à une protéine) aux tests *in vivo*. Au final, il est toujours préférable d'identifier des ligands à la fois bio-actifs et 'drug-ables'.

J'ai évoqué, précédemment, les avantages, en terme de coût, de temps et d'infrastructure nécessaire, du criblage virtuel comme technique support au criblage expérimental ; constat encore plus évident dans un contexte de travail académique. Le criblage virtuel permet d'explorer un espace des molécules très important afin d'en extraire un sous-ensemble qui contiendra les ligands potentiels. Les succès obtenus montrent que cette technique est suffisamment mature pour être associées aux criblages expérimentaux.

La détermination des propriétés associées aux drogues a fait l'objet de nombreux travaux '(drug-likeness' (Clark and Pickett, 2000; Muegge, 2003; Veber, et al., 2002; Vieth, et al., 2004)). Les modèles statistiques ont montré qu'historiquement les drogues absorbées oralement possèdent les 5 règles de Lipinski (Lipinski, et al., 1997). A cela on peut ajouter un nombre maximum de rotamères (7) et une surface polaire maximale de 140 $Å^2$ (Veber, et al., 2002). Plus récemment, le concept de 'lead-like' a été créé pour caractériser les ligands issus des criblages 'haut débit' expérimentaux (HTS) mais qui doivent ensuite être optimisés en drogues. Cette étape n'est cependant pas aussi aisée car on ne sait toujours pas modéliser ni prédire précisément ce qu'est une drogue.

Le criblage de fragments se pose en alternative aux criblages de chimiothèques classiques dont l'objectif est d'identifier des 'leads' (Lepre, 2001). Les fragments sont des molécules commerciales de petit poids moléculaire (<300). Il a été montré que des petites molécules polaires sont plus efficaces pour identifier des touches lors des criblages car elles sont diverses (leur petit poids moléculaire et donc leur plus petit nombre permet d'explorer un espace chimique relatif plus grand), affines (les fragments sont suffisamment petits pour éviter les interactions défavorables), optimisables car pas encore trop complexes et gérables au sein d'un laboratoire grâce à la petite taille de la chimiothèque (stockage par exemple).

L'utilisation de fragments pour identifier de nouveaux candidats-médicaments dans l'industrie pharmaceutique a conduit, avec succès, à l'identification de nouvelles têtes de séries pour plusieurs cibles protéiques (Nienaber, et al., 2000). Cette stratégie combine la chémoinformatique, les criblages structuraux expérimentaux (petit débit) et la chimie médicinale. L'intérêt du criblage de fragments est lié à l'utilisation des techniques comme la RMN et/ou la cristallographie (Hartshorn, et al., 2005; Lepre, et al., 2002; Nienaber, et al., 2000; Sharff and Jhoti, 2003) :

- Les techniques expérimentales de type RMN et cristallographie permettent de connaître le mode de fixation du ou des fragments dans le site actif de la protéine même si l'affinité est faible (jusqu'au millimolaire). Ces ligands passeraient inaperçus dans des tests expérimentaux *in vitro* ou *in vivo*.

- L'information donnée par la fixation, même faible, d'un fragment peut être utilisée pour dériver des molécules de plus haut poids moléculaire plus affines.
- Les fragments sont peu fonctionnalisés chimiquement. On suppose alors qu'ils ont une capacité plus grande à se fixer dans un site actif sans créer de gêne stérique. La contrepartie est qu'ils sont moins spécifiques et qu'ils peuvent se fixer dans différents sites de la protéine. Cette difficulté sera levée lors de l'étape de grossissement de la molécule.
- Si les fragments sont apparentés à ceux que l'on trouve dans les drogues ou les molécules 'bio', alors ils possèderont plus probablement des propriétés 'drug-like'. Cette catégorie de fragments apparentés aux drogues permet de limiter l'espace des fragments à cribler [500-1000]. Ce nombre est très faible face au 250 000 molécules de type 'lead' qu'il faut en moyenne cribler expérimentalement pour espérer identifier une touche intéressante dans des tests HTS *in vitro* (Hibert and Haiech, 2000).
- Enfin, l'avantage des fragments sur des molécules de type 'lead' est qu'ils seront plus facilement modifiables/orientables lors de l'étape d'optimisation/maturation afin de posséder des propriétés plus 'drug-ables'.

Cette approche par fragment n'est pas sans rappeler le concept de combinaison de fragments utilisé par mon programme de *de novo* 'drug design' LEA3D. La manipulation des fragments moléculaires ainsi que leur combinaison s'appuieront donc naturellement sur cette expérience. Un certain nombre de cas tests sont déjà en cours d'étude en collaboration avec des expérimentalistes du Centre de Biochimie Structural de Montpellier (J.-F. Guichou (MCF), M. Cohen-Gonsaud (CNRS) et P. Barthe (IR2-Université)).

Ma contribution à ce projet a tout d'abord consisté à élaborer une chimiothèque de fragments commerciaux. Deux bibliothèques de 10287 et 912 molécules ont été réalisées. La première contient des fragments commerciaux aux propriétés bien définies dont celle de faire partie des familles structurales observées dans des drogues déjà commercialisées. La seconde base contient des fragments de drogues que j'ai pu retrouver tels quels dans des bases de molécules commerciales ; Ils portent le pharmacophore partiel voire total de la drogue à laquelle ils appartiennent. Ces deux chimiothèques de fragments seront utilisées par criblage virtuel afin de ne retenir qu'un petit nombre de molécules qui sera ensuite criblé expérimentalement par RMN et/ou en co-cristallisation avec la protéine purifiée. On bénéficie de l'expertise d'un chimiste pour affiner le choix des molécules (en terme de réactivité et de potentiel de dérivation de la molécule). La base physique des fragments au laboratoire se constitue peu à peu à partir des molécules sélectionnées et achetées.

Les premiers criblages sur la Cyclophilin humaine de type D (étudiées pour leur implication dans HIV (Yoo, et al., 1997), HCV (Watashi and Shimotohno, 2007) et en oncologie (Yao, et al., 2005)) ont permis de d'identifier, par RMN, 4 ligands parmi les 30 sélectionnés par criblage virtuel. 3 d'entre eux ont ensuite été co-cristallisés par trempage avec la protéine. Ils présentent différents modes de fixation (figure 20). La localisation et la description atomique détaillée de l'interaction entre les fragments et la protéine vont permettre dans un second temps de guider la recherche vers des composés plus élaborés.

A ce stade de l'étude, la chémoinformatique intervient à nouveau dans la maturation/optimisation des touches en ligands plus élaborés et plus affins. Trois stratégies sont employées : lier deux fragments par un connecteur (suggestions proposées par *de novo* 'drug design' par LEA3D) ou bien grossir une molécule par ajout d'un autre fragment (par exemple, via la sélection de réactifs : cette molécule est accessible par synthèse) ou bien de chercher des analogues structuraux commerciaux disponibles rapidement (présentant un

pharmacophore ou une sous-structure similaire). Les stratégies basées sur la sélection de molécules commerciales seront privilégiées afin d'éviter une synthèse chimique (Figure 21).

Ce type de criblage est envisagé pour les Cyclophilines humaines (Gothel and Marahiel, 1999; Guichou, et al., 2006) (A, B et D), Pin1 (mitotic rotamase (Lu, 2004)) et les RPFs (resuscitation promoting factors (Cohen-Gonsaud, et al., 2004)) de *Mycobacterium tuberculosis*, protéines produites au laboratoires.

D'autres collaborations sont ou seront établies pour effectuer les tests *in vitro* et *in vivo* ainsi que pour d'éventuelles étapes de synthèse chimique.

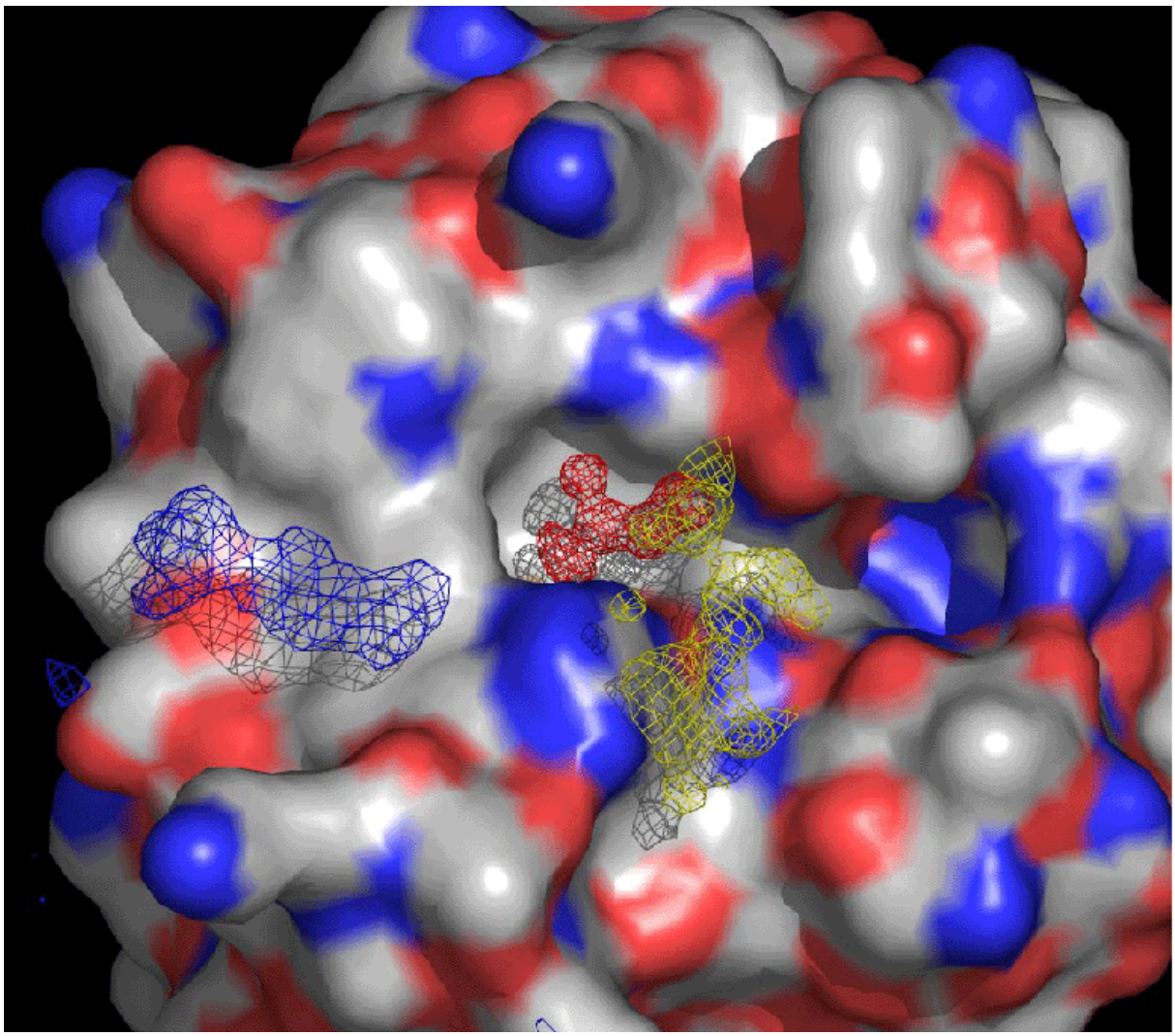


**Figure 20** : Superposition de 3 structures cristallographiques de la cyclophiline D centrées sur le site actif et liées à 3 fragments différents représentés par leur densité électronique. La résolution varie de 1 à 1.4 Å.

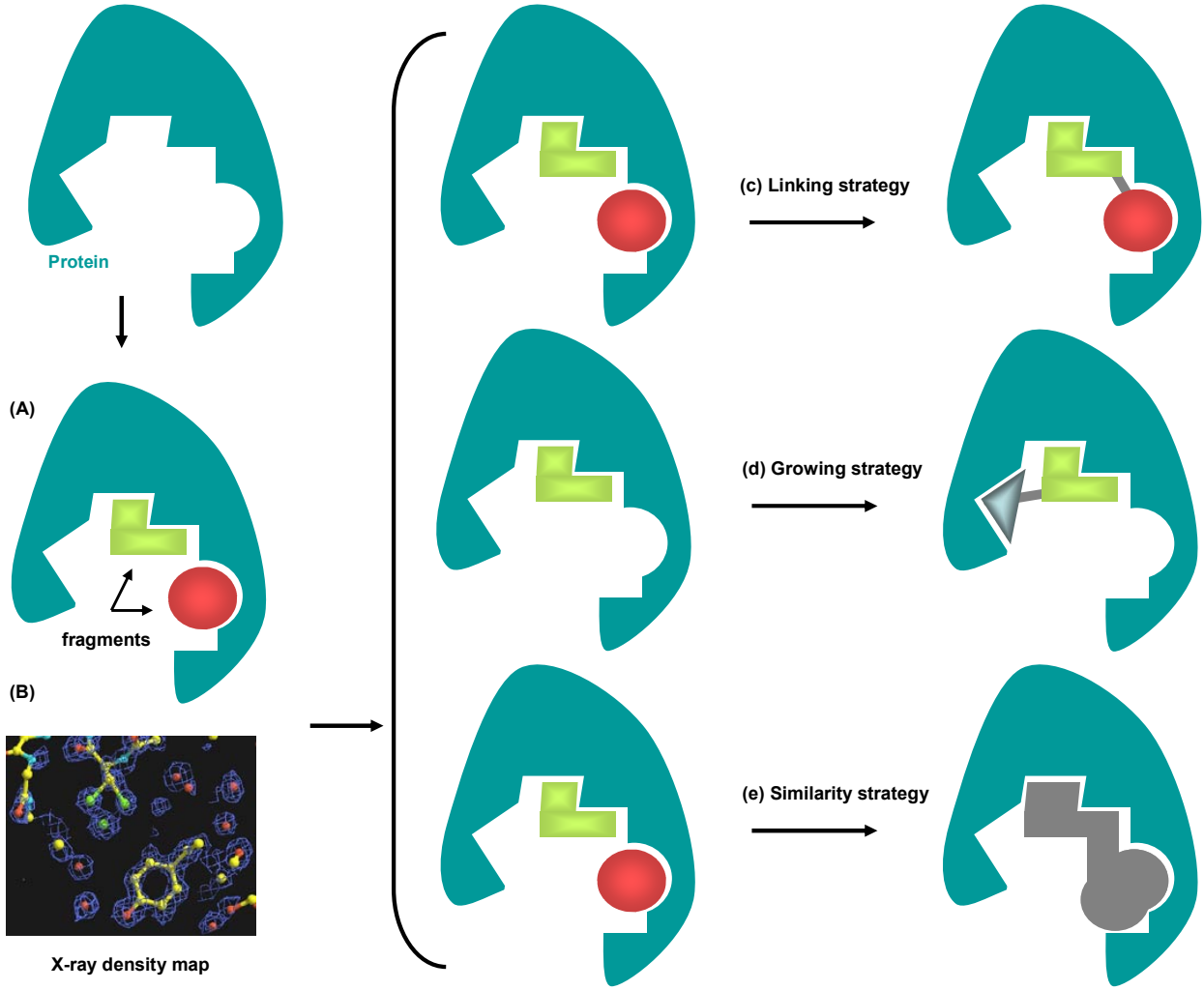



**Figure 21** : Identification de molécules bio-actives basée sur le criblage de fragments.
(a) Les molécules se fixant dans le site actif de la protéine peuvent servir d'amorces à la conception de molécules plus élaborées et plus affines.
(b) La densité électronique issue des clichés de diffraction permet de connaître la position exacte des fragments dans le site actif.
Etape de Maturation :
(c) La première méthode de maturation des fragments est la stratégie de connexion par un outil de *de novo* 'drug design'.
(d) La seconde approche consiste à grossir un fragment particulier en une molécule plus importante par un outil de *de novo* 'drug design'.
(e) La troisième stratégie par similarité identifie des molécules commerciales analogues qui contiennent les fragments identifiés comme ligands.

## V - Perspectives

♦ En terme d'avancées dans 'l'état de l'art', deux des trois volets abordés précédemment sont suffisamment matures et fiables pour être prédictifs dans leurs limites d'application. Le moins avancé est la prédiction des interactions entre protéines mais sa marge de progression est par conséquent beaucoup plus intéressante. D'ailleurs, certaines équipes comme celle de David Baker, plutôt connue aux sessions CASPs, ne s'y sont pas trompées et dorénavant elles participent aussi aux sessions CAPRI. En peu de temps, elles se sont hissées au niveau des plus anciens du domaine avec en plus des méthodes innovantes. Un serveur a également fait son apparition avec des résultats probants (ClusPro, http://nrc.bu.edu/cluster/, (Comeau, et al., 2005)) ).

En 2005, le projet DOCKGROUND reçu le soutien du NIH par l'attribution d'un Grant (R01 GM074255-01). Mon travail a donc pu être poursuivi par l'équipe d'Ilya Vakser. Actuellement, la seconde base de données 'unbound' est en cours de réalisation (jeu de données d'environ 500 exemples). Elle devrait permettre de développer de nouvelles approches pour modéliser/prendre en compte les modifications de conformation des protéines lors de la complexation (passage d'une forme libre dans le solvant à une forme complexée ('induce-fit')). Ces approches sont indispensables si l'on veut travailler à partir de modèles de protéines. Il est à noté qu'il existe un jeu de données semblable réalisé manuellement mais contenant moins d'une centaine d'exemples (hétéro-complexes dont chaque entité a été déterminée expérimentalement d'une façon isolée (Mintseris, et al., 2005)).

Un autre aspect de ce travail n'a pas encore été abordé à ma connaissance. Il s'agit des complexes dont l'interface inclut un ligand. Une analyse préliminaire de notre base de données fait état de plus de 600 cas d'homodimères pour lesquels 3 résidus au moins de chaque chaîne interagit avec le ligand identifié après exclusion des ions, solvants et acides aminés modifiés. Parallèlement, seulement 4 cas d'hétérodimères sont répertoriés dont le plus connus ARF1-GDP/Sec7/Bréfeldine (PDB1re0)(Pommier and Cherfils, 2005).

Il serait intéressant d'analyser plus précisément ces complexes ternaires biologiques ou même artificiels (cristallographiques) pour comprendre le rôle du ligand dans l'interface. On peut espérer identifier quelques cas d'études qui pourraient faire l'objet d'expériences de mutagenèse dirigée et/ou de mesures car calorimétrie. La caractérisation d'une interface telle que le présente G. Schreiber pourrait s'appliquer ici d'une façon très instructive car comme je l'ai mentionné précédemment ils corrèlent l'affinité d'un complexe à l'organisation interne de l'interface en terme de nombre et de densité des connectivités entre clusters de résidus (Reichmann, et al., 2005). Cette représentation peut aussi expliquer pourquoi la mutation de certains résidus inhibe toute interaction. L'autre perspective d'étude plus classique serait de comparer le site actif créé par l'association de deux protéines (deux demi site à la surface des protéines) avec les propriétés des sites actifs habituellement localisés au cœur des protéines.

♦ La modélisation par homologie et la prédiction des interactions entre protéines sont des domaines d'études essentiellement académiques tant en méthodologie qu'en application contrairement à la chémoinformatique (criblage virtuel, *de novo* 'drug design', QSAR, …). Cela s'explique par l'ancienneté du champ d'étude (et certainement par l'intérêt) mais aussi par l'accès aux données sources, publiques pour les protéines (PDB) et majoritairement propriétaires pour la caractérisation des molécules. L'investissement dans le haut débit et par conséquent dans la production de données et le 'data mining' est souvent réservé aux industriels. Par exemple, les criblages basés sur les fragments ont émergés au travers d'entreprises de biotechnologie comme Astex ou Vertex Pharmaceuticals. Ils sont maintenant repris par les entreprises pharmaceutiques mais aussi par des académiques. D'ailleurs, le

projet d'identification de molécules bio-actives basée sur le criblage de fragments que j'ai exposé au chapitre précédent vient d'être sélectionné pour un financement de 36 mois par l'ANR dans la catégorie Jeunes Chercheurs / Jeunes Chercheuses.

L'apport de la chémoinformatique académique n'est pourtant pas négligeable puisque la majorité des entreprises pharmaceutiques utilisent des logiciels commerciaux créés à l'origine par des académiques et qui ont été professionnalisés pour en améliorer l'utilisation (système d'exploitation, interface utilisateur, interface avec d'autres programmes…). Le programme LEA3D décrit au chapitre IV fait actuellement l'objet d'un transfert de technologie avec la société Nova Decision (http://www.novadecision.com/) en incubation au Centre de Biochimie Structural à Montpellier (projet d'entreprise de Mr Yasri). Ce projet est soutenu par l'INSERM et le LRI (Languedoc-Roussillon Incubation). Le projet a aussi remporté cette année le concours de la 9$^{ième}$ édition d'aide à la création d'entreprises de technologies innovantes organisé par le ministère de la recherche. A partir de l'automne prochain, je serais davantage impliquée dans cette société à laquelle j'apporterais mon concours scientifique.

Cette habilitation à dirigée des recherches s'effectue au moment même où je m'apprête à changer d'affection en faveur de l'Institut de Pharmacologie Moléculaire et Cellulaire (IPMC) de Sophia-Antipolis afin de permettre un regroupement familial. Mes sujets d'étude vont donc progressivement converger vers les centres d'intérêt de l'IPMC. Déjà, une collaboration est en cours avec le Dr. G. Lambeau sur l'identification de ligands de certaines phospholipases $A_2$ sécrétées humaines pour lesquelles il n'y a pas de ligands connus (cas de la hGIII donnée en exemple dans la section II) ou bien pour lesquelles on souhaite diversifier les familles structurales de ligands ou encore améliorer la spécificité.

## Références


Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M.J., Dumontier, M.R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J.P., Parker, B., Pintilie, G., Pirone, R., Salama, J.J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B.F. and Hogue, C.W. (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Res*, **33**, D418-424.

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.C., Bork, P., Superti-Furga, G., Serrano, L. and Russell, R.B. (2004) Structure-based assembly of protein complexes in yeast, *Science*, **303**, 2026-2029.

Aloy, P., Ceulemans, H., Stark, A. and Russell, R.B. (2003) The relationship between sequence and interaction divergence in proteins, *J Mol Biol*, **332**, 989-998.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics, *Science*, **294**, 93-96.

Bates, P.A., Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM, *Proteins*, **Suppl 5**, 39-46.

Baurin, N., Baker, R., Richardson, C., Chen, I., Foloppe, N., Potter, A., Jordan, A., Roughley, S., Parratt, M., Greaney, P., Morley, D. and Hubbard, R.E. (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds, *J Chem Inf Comput Sci*, **44**, 643-651.

Boehm, H.J., Boehringer, M., Bur, D., Gmuender, H., Huber, W., Klaus, W., Kostrewa, D., Kuehne, H., Luebbers, T., Meunier-Keller, N. and Mueller, F. (2000) Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening, *J Med Chem*, **43**, 2664-2674.

Bogan, A.A. and Thorn, K.S. (1998) Anatomy of hot spots in protein interfaces, *J Mol Biol*, **280**, 1-9.

Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure, *Science*, **253**, 164-170.

Brenner, S.E., Koehl, P. and Levitt, M. (2000) The ASTRAL compendium for protein structure and sequence analysis, *Nucleic Acids Res*, **28**, 254-256.

Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Structure prediction meta server, *Bioinformatics*, **17**, 750-751.

Carugo, O. and Argos, P. (1997) Protein-protein crystal-packing contacts, *Protein Sci*, **6**, 2261-2263.

Catherinot, V. and Labesse, G. (2004) ViTO: tool for refinement of protein sequence-structure alignments, *Bioinformatics*, **20**, 3694-3696.

Cazals, F., Proust, F., Bahadur, R.P. and Janin, J. (2006) Revisiting the Voronoi description of protein-protein interfaces, *Protein Sci*, **15**, 2082-2092.

Chiche, L., Gregoret, L.M., Cohen, F.E. and Kollman, P.A. (1990) Protein model structure evaluation using the solvation free energy of folding, *Proc Natl Acad Sci U S A*, **87**, 3240-3243.

Chivian, D., Kim, D.E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C.E., Bonneau, R., Rohl, C.A. and Baker, D. (2003) Automated prediction of CASP-5 structures using the Robetta server, *Proteins*, **53 Suppl 6**, 524-533.

Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins, *Embo J*, **5**, 823-826.

Clark, D.E. and Pickett, S.D. (2000) Computational methods for the prediction of 'drug-likeness', *Drug Discov Today*, **5**, 49-58.

Cohen-Gonsaud, M., Keep, N.H., Davies, A.P., Ward, J., Henderson, B. and Labesse, G. (2004) Resuscitation-promoting factors possess a lysozyme-like domain, *Trends Biochem Sci*, **29**, 7-10.

Colovos, C. and Yeates, T.O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions, *Protein Sci*, **2**, 1511-1519.

Combet, C., Blanchet, C., Geourjon, C. and Deleage, G. (2000) NPS@: network protein sequence analysis, *Trends Biochem Sci*, **25**, 147-150.

Combet, C., Jambon, M., Deleage, G. and Geourjon, C. (2002) Geno3D: automatic comparative molecular modelling of protein, *Bioinformatics*, **18**, 213-214.

Comeau, S.R., Vajda, S. and Camacho, C.J. (2005) Performance of the first protein docking server ClusPro in CAPRI rounds 3-5, *Proteins*, **60**, 239-244.

Cuff, J.A. and Barton, G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins*, **40**, 502-511.

Dasgupta, S., Iyer, G.H., Bryant, S.H., Lawrence, C.E. and Bell, J.A. (1997) Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers, *Proteins*, **28**, 494-514.

Davis, F.P., Braberg, H., Shen, M.Y., Pieper, U., Sali, A. and Madhusudhan, M.S. (2006) Protein complex compositions predicted by structural similarity, *Nucleic Acids Res*, **34**, 2943-2952.

Davis, F.P. and Sali, A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces, *Bioinformatics*, **21**, 1901-1907.

Davis, I.W., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes, *Nucleic Acids Res*, **32**, W615-619.

Devos, D. and Valencia, A. (2000) Practical limits of function prediction, *Proteins*, **41**, 98-107.

Doman, T.N., McGovern, S.L., Witherbee, B.J., Kasten, T.P., Kurumbail, R., Stallings, W.C., Connolly, D.T. and Shoichet, B.K. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B, *J Med Chem*, **45**, 2213-2221.

Douguet, D., Bolla, J.-M., Munier-Lehmann, H. and Labesse, G. (2002) From sequence to structure to function: a case study, *Enzyme and Microbial Technology*, **30**, 289-294.

Douguet, D., Chen, H.C., Tovchigrechko, A. and Vakser, I.A. (2006) DOCKGROUND resource for studying protein-protein interfaces, *Bioinformatics*, **22**, 2612-2618.

Douguet, D. and Labesse, G. (2001) Easier threading through web-based comparisons and cross-validations, *Bioinformatics*, **17**, 752-753.

Douguet, D., Munier-Lehmann, H., Labesse, G. and Pochet, S. (2005) LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design, *J Med Chem*, **48**, 2457-2468.

Douguet, D., Thoreau, E. and Grassy, G. (1999) A Quantitative Structure-Activity Relationships Studies of RAR a, b, g Retinoid Agonists, *Quant. Struct.-Act. Relat.*, **18**, 107-123.

Douguet, D., Thoreau, E. and Grassy, G. (2000) A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm, *J Comput Aided Mol Des*, **14**, 449-466.

Dunbrack, R.L., Jr. (1999) Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL, *Proteins*, **Suppl 3**, 81-87.

Eisenhaber, F. and Argos, P. (1993) Improved strategy in analytic surface calculation for molecular systems: Handling singularities and computational efficiency, *J. Computat. Chem.*, **11**, 1272-1280.

Finn, R.D., Marshall, M. and Bateman, A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions, *Bioinformatics*, **21**, 410-412.

Fioravanti, E., Haouz, A., Ursby, T., Munier-Lehmann, H., Delarue, M. and Bourgeois, D. (2003) Mycobacterium tuberculosis thymidylate kinase: structural studies of intermediates along the reaction pathway, *J Mol Biol*, **327**, 1077-1092.

Fischer, D., Rychlewski, L., Dunbrack, R.L., Jr., Ortiz, A.R. and Elofsson, A. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods, *Proteins*, **53 Suppl 6**, 503-516.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery, *Nature*, **440**, 631-636.

Gelly, J.C., Chiche, L. and Gracy, J. (2005) EvDTree: structure-dependent substitution profiles based on decision tree classification of 3D environments, *BMC Bioinformatics*, **6**, 4.

Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Bolser, D.M., Oh, D., Kim, D.S. and Bhak, J. (2005) A protein domain interaction interface database: InterPare, *BMC Bioinformatics*, **6**, 207.

Gothel, S.F. and Marahiel, M.A. (1999) Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts, *Cell Mol Life Sci*, **55**, 423-436.

Gracy, J., Chiche, L. and Sallantin, J. (1993) Learning and alignment methods applied to protein structure prediction, *Biochimie*, **75**, 353-361.

Gray, J.J. (2006) High-resolution protein-protein docking, *Curr Opin Struct Biol*, **16**, 183-193.

Gruneberg, S., Stubbs, M.T. and Klebe, G. (2002) Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation, *J Med Chem*, **45**, 3588-3602.

Guichou, J.F., Viaud, J., Mettling, C., Subra, G., Lin, Y.L. and Chavanieu, A. (2006) Structure-based design, synthesis, and biological evaluation of novel inhibitors of human cyclophilin A, *J Med Chem*, **49**, 900-910.

Hartshorn, M.J., Murray, C.W., Cleasby, A., Frederickson, M., Tickle, I.J. and Jhoti, H. (2005) Fragment-based lead discovery using X-ray crystallography, *J Med Chem*, **48**, 403-413.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M.J. (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force, *J Mol Biol*, **216**, 167-180.

Henrick, K. and Thornton, J.M. (1998) PQS: a protein quaternary structure file server, *Trends Biochem Sci*, **23**, 358-361.

Hibert, M. and Haiech, J. (2000) Des gènes aux médicaments : nouveaux défis, nouvelles stratégies, *médecine/sciences*, **16**, 1332-1339.

Hubbard, T.J., Murzin, A.G., Brenner, S.E. and Chothia, C. (1997) SCOP: a structural classification of proteins database, *Nucleic Acids Res*, **25**, 236-239.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I. and Wodak, S.J. (2003) CAPRI: a Critical Assessment of PRedicted Interactions, *Proteins*, **52**, 2-9.

Janin, J. and Rodier, F. (1995) Protein-protein interaction at crystal contacts, *Proteins*, **23**, 580-587.

Jefferson, E.R., Walsh, T.P. and Barton, G.J. (2006) Biological units and their effect upon the properties and prediction of protein-protein interactions, *J Mol Biol*, **364**, 1118-1129.

Jones, D.T., Tress, M., Bryson, K. and Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure, *Proteins*, **Suppl 3**, 104-111.

Jones, S. and Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc Natl Acad Sci U S A*, **93**, 13-20.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, **14**, 846-856.

Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J Mol Biol*, **299**, 499-520.

Keskin, O., Bahar, I., Badretdinov, A.Y., Ptitsyn, O.B. and Jernigan, R.L. (1998) Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions, *Protein Sci*, **7**, 2578-2586.

Keskin, O., Tsai, C.J., Wolfson, H. and Nussinov, R. (2004) A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications, *Protein Sci*, **13**, 1043-1055.

Korkin, D., Davis, F.P., Alber, F., Luong, T., Shen, M.Y., Lucic, V., Kennedy, M.B. and Sali, A. (2006) Structural modeling of protein interactions by analogy: application to PSD-95, *PLoS Comput Biol*, **2**, e153.

Kryshtafovych, A., Venclovas, C., Fidelis, K. and Moult, J. (2005) Progress over the first decade of CASP experiments, *Proteins*, **61 Suppl 7**, 225-236.

Labesse, G., Douguet, D., Assairi, L. and Gilles, A.M. (2002) Diacylglyceride kinases, sphingosine kinases and NAD kinases: distant relatives of 6-phosphofructokinases, *Trends Biochem Sci*, **27**, 273-275.

Labesse, G. and Mornon, J. (1998) Incremental threading optimization (TITO) to help alignment and modelling of remote homologues, *Bioinformatics*, **14**, 206-211.

Larsen, T.A., Olson, A.J. and Goodsell, D.S. (1998) Morphology of protein-protein interfaces, *Structure*, **6**, 421-427.

Lepre, C.A. (2001) Library design for NMR-based screening, *Drug Discov Today*, **6**, 133-140.

Lepre, C.A., Peng, J., Fejzo, J., Abdul-Manan, N., Pocas, J., Jacobs, M., Xie, X. and Moore, J.M. (2002) Applications of SHAPES screening in drug discovery, *Comb Chem High Throughput Screen*, **5**, 583-590.

Levy, E.D., Pereira-Leal, J.B., Chothia, C. and Teichmann, S.A. (2006) 3D complex: a structural classification of protein complexes, *PLoS Comput Biol*, **2**, e155.

Li de la Sierra, I., Munier-Lehmann, H., Gilles, A.M., Barzu, O. and Delarue, M. (2001) X-ray structure of TMP kinase from Mycobacterium tuberculosis complexed with TMP at 1.95 A resolution, *J Mol Biol*, **311**, 87-100.

Lichtarge, O., Bourne, H.R. and Cohen, F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol*, **257**, 342-358.

Lijnzaad, P. and Argos, P. (1997) Hydrophobic patches on protein subunit interfaces: characteristics and prediction, *Proteins*, **28**, 333-343.

Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) Experimental and computational approaches to estmate solubility and permeability in drug discovery and development settings., *Adv. Drug Delivery Rev.*, **23**, 4-25.

Lo Conte, L., Chothia, C. and Janin, J. (1999) The atomic structure of protein-protein recognition sites, *J Mol Biol*, **285**, 2177-2198.

Lu, H., Lu, L. and Skolnick, J. (2003) Development of unified statistical potentials describing protein-protein interactions, *Biophys J*, **84**, 1895-1901.

Lu, K.P. (2004) Pinning down cell signaling, cancer and Alzheimer's disease, *Trends Biochem Sci*, **29**, 200-209.

Lu, L., Arakaki, A.K., Lu, H. and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome, *Genome Res*, **13**, 1146-1154.

Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R. (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces, *Proc Natl Acad Sci U S A*, **100**, 5772-5777.

Mendez, R., Leplae, R., Lensink, M.F. and Wodak, S.J. (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures, *Proteins*, **60**, 150-169.

Michalickova, K., Bader, G.D., Dumontier, M., Lieu, H., Betel, D., Isserlin, R. and Hogue, C.W. (2002) SeqHound: biological sequence and structure database as a platform for bioinformatics research, *BMC Bioinformatics*, **3**, 32.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. and Weng, Z. (2005) Protein-Protein Docking Benchmark 2.0: an update, *Proteins*, **60**, 214-216.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV, *Proteins*, **Suppl 5**, 2-7.

Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V, *Proteins*, **53 Suppl 6**, 334-339.

Muegge, I. (2003) Selection criteria for drug-like compounds, *Med Res Rev*, **23**, 302-321.

Nienaber, V.L., Richardson, P.L., Klighofer, V., Bouska, J.J., Giranda, V.L. and Greer, J. (2000) Discovering novel ligands for macromolecules using X-ray crystallographic screening, *Nat Biotechnol*, **18**, 1105-1108.

Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., Shen, M.Y., Kelly, L., Melo, F. and Sali, A. (2006) MODBASE: a database of annotated comparative protein structure models and associated resources, *Nucleic Acids Res*, **34**, D291-295.

Pochet, S., Dugue, L., Douguet, D., Labesse, G. and Munier-Lehmann, H. (2002) Nucleoside analogues as inhibitors of thymidylate kinases: possible therapeutic applications, *Chembiochem*, **3**, 108-110.

Pommier, Y. and Cherfils, J. (2005) Interfacial inhibition of macromolecular interactions: nature's paradigm for drug discovery, *Trends Pharmacol Sci*, **26**, 138-145.

Powers, R.A., Morandi, F. and Shoichet, B.K. (2002) Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase, *Structure*, **10**, 1013-1023.

Reichmann, D., Rahat, O., Albeck, S., Meged, R., Dym, O. and Schreiber, G. (2005) The modular architecture of protein-protein binding interfaces, *Proc Natl Acad Sci U S A*, **102**, 57-62.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T. and Vriend, G. (1998) Homology modeling, model and software evaluation: three related resources, *Bioinformatics*, **14**, 523-528.

Russ, A.P. and Lampel, S. (2005) The druggable genome: an update, *Drug Discov Today*, **10**, 1607-1610.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M. and Sali, A. (2004) A structural perspective on protein-protein interactions, *Curr Opin Struct Biol*, **14**, 313-324.

Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints, *J Mol Biol*, **234**, 779-815.

Schneider, G. and Bohm, H.J. (2002) Virtual screening and fast automated docking methods, *Drug Discov Today*, **7**, 64-70.

Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: An automated protein homology-modeling server, *Nucleic Acids Res*, **31**, 3381-3385.

Sharff, A. and Jhoti, H. (2003) High-throughput crystallography to enhance drug discovery, *Curr Opin Chem Biol*, **7**, 340-345.

Shi, J., Blundell, T.L. and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, *J Mol Biol*, **310**, 243-257.

Shoichet, B.K. (2004) Virtual screening of chemical libraries, *Nature*, **432**, 862-865.

Stein, A., Russell, R.B. and Aloy, P. (2005) 3did: interacting protein domains of known three-dimensional structure, *Nucleic Acids Res*, **33**, D413-417.

Strynadka, N.C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B.K., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. and James, M.N. (1996) Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase, *Nat Struct Biol*, **3**, 233-239.

Teyra, J., Doms, A., Schroeder, M. and Pisabarro, M.T. (2006) SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces, *BMC Bioinformatics*, **7**, 104.

Todd, A.E., Marsden, R.L., Thornton, J.M. and Orengo, C.A. (2005) Progress of structural genomics initiatives: an analysis of solved target structures, *J Mol Biol*, **348**, 1235-1260.

Topham, C.M., Thomas, P., Overington, J.P., Johnson, M.S., Eisenmenger, F. and Blundell, T.L. (1990) An assessment of COMPOSER: a rule-based approach to modelling protein structure, *Biochem Soc Symp*, **57**, 1-9.

Torda, A.E. (1997) Perspectives in protein-fold recognition, *Curr Opin Struct Biol*, **7**, 200-205.

Tress, M., Ezkurdia, I., Grana, O., Lopez, G. and Valencia, A. (2005) Assessment of predictions submitted for the CASP6 comparative modeling category, *Proteins*, **61 Suppl 7**, 27-45.

Tsai, C.J., Lin, S.L., Wolfson, H.J. and Nussinov, R. (1996) A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique, *J Mol Biol*, **260**, 604-620.

Vajda, S., Vakser, I.A., Sternberg, M.J. and Janin, J. (2002) Modeling of protein interactions in genomes, *Proteins*, **47**, 444-446.

Vanheusden, V., Munier-Lehmann, H., Froeyen, M., Busson, R., Rozenski, J., Herdewijn, P. and Van Calenbergh, S. (2004) Discovery of bicyclic thymidine analogues as selective and high-affinity inhibitors of Mycobacterium tuberculosis thymidine monophosphate kinase, *J Med Chem*, **47**, 6187-6194.

Veber, D.F., Johnson, S.R., Cheng, H.Y., Smith, B.R., Ward, K.W. and Kopple, K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates, *J Med Chem*, **45**, 2615-2623.

Velyvis, A., Vaynberg, J., Yang, Y., Vinogradova, O., Zhang, Y., Wu, C. and Qin, J. (2003) Structural and functional insights into PINCH LIM4 domain-mediated integrin signaling, *Nat Struct Biol*, **10**, 558-564.

Venclovas, C., Zemla, A., Fidelis, K. and Moult, J. (2001) Comparison of performance in successive CASP experiments, *Proteins*, **Suppl 5**, 163-170.

Venclovas, C., Zemla, A., Fidelis, K. and Moult, J. (2003) Assessment of progress over the CASP experiments, *Proteins*, **53 Suppl 6**, 585-595.

Vieth, M., Siegel, M.G., Higgs, R.E., Watson, I.A., Robertson, D.H., Savin, K.A., Durst, G.L. and Hipskind, P.A. (2004) Characteristic physical properties and structural fragments of marketed oral drugs, *J Med Chem*, **47**, 224-232.

von Itzstein, M., Wu, W.Y., Kok, G.B., Pegg, M.S., Dyason, J.C., Jin, B., Van Phan, T., Smythe, M.L., White, H.F., Oliver, S.W. and et al. (1993) Rational design of potent sialidase-based inhibitors of influenza virus replication, *Nature*, **363**, 418-423.

Wallner, B. and Elofsson, A. (2003) Can correct protein models be identified?, *Protein Sci*, **12**, 1073-1086.

Walters, W.P., Stahl, M.T. and Murcko, M.A. (1998) Virtual screening - an overview, *Drug Discovery Today*, **3**, 160-178.

Wang, G. and Dunbrack, R.L., Jr. (2003) PISCES: a protein sequence culling server, *Bioinformatics*, **19**, 1589-1591.

Watashi, K. and Shimotohno, K. (2007) Chemical genetics approach to hepatitis C virus replication: cyclophilin as a target for anti-hepatitis C virus strategy, *Rev Med Virol*, **17**, 245-252.

Winter, C., Henschel, A., Kim, W.K. and Schroeder, M. (2006) SCOPPI: a structural classification of protein-protein interfaces, *Nucleic Acids Res*, **34**, D310-314.

Wlodawer, A. and Vondrasek, J. (1998) Inhibitors of HIV-1 protease: a major success of structure-assisted drug design, *Annu Rev Biophys Biomol Struct*, **27**, 249-284.

Xu, Q., Canutescu, A., Obradovic, Z. and Dunbrack, R.L., Jr. (2006) ProtBuD: a database of biological unit structures of protein families and superfamilies, *Bioinformatics*, **22**, 2876-2882.

Yao, Q., Li, M., Yang, H., Chai, H., Fisher, W. and Chen, C. (2005) Roles of cyclophilins in cancers and other organ systems, *World J Surg*, **29**, 276-280.

Yoo, S., Myszka, D.G., Yeh, C., McMurray, M., Hill, C.P. and Sundquist, W.I. (1997) Molecular recognition in the HIV-1 capsid/cyclophilin A complex, *J Mol Biol*, **269**, 780-795.

Zhang, Z., Schaffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V. and Altschul, S.F. (1998) Protein sequence similarity searches using patterns as seeds, *Nucleic Acids Res*, **26**, 3986-3990.

Zhou, Z.H., Baker, M.L., Jiang, W., Dougherty, M., Jakana, J., Dong, G., Lu, G. and Chiu, W. (2001) Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus, *Nat Struct Biol*, **8**, 868-873.

## Articles